

Inverse Optimization with Noisy Data

Anil Aswani, Zuo-Jun Max Shen, Auyon Siddiq

Department of Industrial Engineering and Operations Research, University of California, Berkeley, USA,
aaswani@berkeley.edu, maxshen@berkeley.edu, auyon.siddiq@berkeley.edu.

Inverse optimization consists of determining unknown parameters of an optimization problem based on knowledge of its optimal points. This paper considers inverse optimization in the setting where measurements of the optimal points of a convex optimization problem are corrupted by noise. We first provide a formulation for the inverse optimization problem with noisy data and show it is NP-hard, and then we prove that existing convex optimization-based heuristics for inverse optimization with noisy data are statistically inconsistent (meaning the answers provided by these methods generally converge to the “wrong” answer as more data is collected). Next, we show that our formulation is statistically consistent by combining a new duality-based reformulation for bilevel programs with a regularization scheme that smooths discontinuities in our formulation. Using epi-convergence theory, we show the regularization parameter can be adjusted to approximate the original inverse optimization problem to arbitrary accuracy, and this is used to prove our statistical consistency results. This duality-based reformulation is next used to propose two numerical algorithms for solving the inverse optimization problem with noisy data. The first is an enumeration algorithm that is applicable to settings where the dimensionality of the parameter space is modest, and the second is a semiparametric approach that combines nonparametric statistics with a modified version of our formulation of the inverse optimization problem. These numerical algorithms are shown to maintain the statistical consistency of our formulation. Lastly, we demonstrate using synthetic and real data sets the competitiveness of our numerical algorithms as compared to existing heuristics.

Key words: statistics: estimation; programming: nonlinear; utility/preference: estimation

1. Introduction

An appreciable share of real-world data represents *decisions*, which can often be characterized as the solutions of correspondingly-defined optimization problems. Estimating the parameters of these latent optimization problems has the potential to provide greater insight into how decisions are made, and also enable the prediction of future decisions. Examples of domains where this is important include health systems engineering (Aswani et al. 2016), energy systems engineering (?), and marketing (?), where such estimation may lead to new approaches that enable the individualization of products and incentives.

For example, consider a single homeowner who *each day* observes an electricity price and weather forecast and then adjusts the temperature set-point for their home’s air-conditioner. By modeling this homeowner’s decision as being generated from an optimization problem, we can directly estimate the price elasticity of comfort – as measured by a standardized function of the temperature set-point and the outside temperature (ASHRAE 2013) – for this particular homeowner.

This information is valuable for designing personalized incentive bonus schemes that encourage participation in demand-response programs (?) or promote energy-efficiency (Aswani and Tomlin 2012).

1.1. Overview

This paper considers the problem of estimating unknown model parameters of an optimization problem based on noisy measurements of optimal solutions of this optimization problem. We broadly call this estimation process: inverse optimization with noisy data. In particular, the novelty of our approach is to provide the first *statistical inference* perspective on the inverse optimization problem. This is important because real-world decision data is noisy, either because (i) the data collection process introduces measurement noise, (ii) the decision-maker deviates from optimal decisions – phenomena often referred to as *bounded rationality* (Tversky and Kahneman 1981), or (iii) there is mismatch between the equations of the model and the actual decision-making process.

Noisy data make inverse optimization challenging because noise in the solution data can preclude the existence of a single set of model parameters that renders all observed solutions exactly optimal. In this setting, the goal of inverse optimization is to find a set of model parameters that achieves a good “fit” with respect to the solution data. More specifically, we are interested in two statistical questions. First, how can we generate estimates of unknown model parameters that asymptotically provide the best possible predictions from the chosen equation for the model? In statistics, this property is known as *risk consistency* (Bartlett and Mendelson 2002, Greenshtein and Ritov 2004, Chatterjee 2014). Second, when the chosen equation for the model matches that which is generating the solution data, how can we generate estimates that asymptotically converge to the true value of the unknown parameters? In statistics, this property is known as *consistency* (Wald 1949, Jennrich 1969, Bickel and Doksum 2006). We will use the term *estimation consistency* to distinguish this concept from risk consistency. Note that estimation consistency generally implies risk consistency.

Restated, a risk consistent estimate asymptotically achieves the lowest possible prediction error (out of all possible predictions permitted by the class of models considered). Hence, risk consistency and estimation consistency allow us to be confident that prediction and estimation accuracy, respectively, will generally improve with additional data. By contrast, an estimator that fails to be risk consistent (so-called *inconsistent* estimators) may yield poor predictions, even if a large amount of data is available. Proving consistency of an estimator is an important topic in the theory of statistical inference (cf. (Wald 1949, Jennrich 1969, Bartlett and Mendelson 2002, Greenshtein and Ritov 2004, Bickel and Doksum 2006, Chatterjee 2014, Aswani 2015)), and consistency is considered to be a minimal requirement for an estimator (Bickel and Doksum 2006).

The main paper begins with Section 2, which describes the statistical and computational challenges of inverse optimization with noisy data. The section begins by formally defining a (convex)

forward optimization problem and its corresponding inverse optimization problem. We specifically formulate the inverse optimization problem such that (as we later show) its solution has the desired statistical consistency properties. Our approach is conceptually similar to least squares regression in the sense that we also employ a sum-of-squares loss function to fit a parametric model to noisy data. The substantive difference is that inverse optimization involves estimating the (possibly multi-valued) solution set of a general convex optimization problem, whereas regression typically involves estimating a (single-valued) function which has a closed form expression. Due to these differences, much of the classical statistical theory on least-squares regression (Jennrich 1969) is invalid in the inverse optimization setting, and thus new analysis is required. We also note that our approach is not restricted to the use of an ℓ_2 norm: Results similar to those in our paper can be proved for other loss functions, such as absolute deviation or a likelihood function, but we do not consider those extensions in this paper.

In Section 3, we study the statistical consistency of our formulation of the inverse optimization problem. The key technical difficulty in proving these results is dealing with continuity issues. In particular, the risk measures are not continuous in the general case, but are rather lower semicontinuous. As alluded to above, this precludes the use of the typical statistical machinery used to prove consistency results (namely the uniform law of large numbers (Jennrich 1969) and related uniform bounds (Bartlett and Mendelson 2002, Greenshtein and Ritov 2004)). To circumvent this difficulty, we define a regularized version of the inverse optimization problem that smooths out any discontinuities, and this regularized version of the problem is constructed using a new duality-based reformulation for bilevel programs. Using epi-convergence theory, we show the regularization parameter can be adjusted to approximate the original inverse optimization problem to arbitrary accuracy. The section concludes by using the regularized version of the inverse optimization problem to prove results on the statistical consistency of our formulation.

Section 4 provides two numerical algorithms for solving our formulation of the inverse optimization problem. The first numerical algorithm is an enumeration algorithm that is applicable to settings where the dimensionality of the parameter space is modest (i.e., at most four or five parameters). The second numerical algorithm is a semiparametric approach that combines nonparametric statistics with a modified version of our formulation of the inverse optimization problem. The statistical consistency of these two numerical algorithms are shown using the results from Section 3. Lastly, in Section 5 we demonstrate using synthetic and real data sets the competitiveness of our approaches as compared to existing heuristics (Keshavarz et al. 2011, Bertsimas et al. 2014).

1.2. Literature Review

Existing inverse optimization models differ based on their specification of the *loss function*, and the different models can be broadly categorized into either (i) deterministic settings, or (ii) noisy

settings. The work in the deterministic setting has primarily focused on single observation situations, wherein a single optimal solution is observed and then used to estimate parameters of the optimization problem. However, in the noisy setting past work has considered situations with either a single observation and multiple observations.

We begin by describing some of the work in the deterministic setting: Ahuja and Orlin (2001) consider the estimation of objective function coefficients of general linear programs given a single optimal solution. The feasible region of the inverse problem is formulated using the constraints of the dual program and complementary slackness conditions. Since the observed solution is assumed to be optimal, feasibility of the inverse problem is guaranteed. Iyengar and Kang (2005) and Zhang and Xu (2010) extend inverse optimization to certain conic forward problems using conic duality theory. Inverse optimization models have also been studied in the context of integer programs (Schaefer 2009, Wang 2009) and network problems (Burton and Toint 1992, Hochbaum 2003, Zhang and Liu 1996). With respect to applications, inverse optimization models has been employed in many different domains, including healthcare (Erkin et al. 2010, Chan et al. 2014), energy (Ratliff et al. 2014, Saez-Gallego et al. 2016), finance (Bertsimas et al. 2012), production planning (Troutt et al. 2006), demand management (Carr and Lovejoy 2000, Bajari et al. 2007), auction design (Beil and Wein 2003), telecommunication (Faragó et al. 2003) and geoscience (Burton and Toint 1992). We refer the reader to Heuberger (2004) for a survey of inverse optimization methods.

The noisy setting has been less studied. Chan et al. (2014) propose a generalized approach to inverse optimization for linear programs where the (single) observed solution may be suboptimal or infeasible. Instead of complementary slackness, the authors use dual feasibility and strong duality to formulate the inverse problem. To accommodate noise, the strong duality constraint is relaxed to guarantee feasibility of the inverse problem. Saez-Gallego et al. (2016) also consider inverse optimization for linear programs, and formulate the inverse problem using KKT conditions. Keshavarz et al. (2011) formulates the inverse problem using the KKT conditions of the optimization problem. To accommodate noise, the KKT conditions are relaxed by introducing slack variables to allow the data to “approximately” satisfy the KKT conditions. Similarly, Bertsimas et al. (2014) consider inverse problems where the observed data are assumed to be in an equilibrium. The authors enforce optimality conditions using a variational inequality, and similarly relax the optimality conditions by introducing slack variables to allow the data to “approximately” satisfy the variational inequality.

Our work in this paper is most closely related to the noisy setting with multiple observations that has been previously considered by Keshavarz et al. (2011) and Bertsimas et al. (2014). The key distinction between our work and these two previous approaches is in the choice of the loss function. In (Keshavarz et al. 2011) and (Bertsimas et al. 2014), the loss function is measured by the amount of slack required to make the measured data satisfy an approximate optimality condition (either the

KKT conditions (Keshavarz et al. 2011) or a variational inequality describing optimality (Bertsimas et al. 2014)). In contrast, our approach is to jointly estimate (i) the parameters of the optimization problem, and (ii) the denoised versions of the measured data (i.e. the true underlying optimal solutions). By performing this joint estimation, we are able to define our loss function to be the average discrepancy between the measured data and the (estimated) denoised data. As we will show, this difference in loss function leads to significantly improved statistical performance. A secondary distinction is that we propose the use of a novel optimality condition: specifically, we upper bound the objective function of a convex optimization problem by its dual – thereby enforcing a zero duality gap and guaranteeing optimality. An important benefit of using this alternate optimality condition is that it has favorable convexity and continuity properties (which are not available when using KKT conditions or variational inequalities to represent optimality) that enable design of numerical algorithms for solving the inverse optimization problem.

1.3. Contributions

Our contributions in this paper include both statistical and optimization results, and there are specifically two main contributions. The first is we show that solving a bilevel formulation for the problem of inverse optimization with noisy data provides parameter estimates that are statistically consistent. This statistical result is independent of the approach used to solve the bilevel formulation. Our second main contribution is to propose two numerical algorithms for solving the bilevel formulation by using a novel duality-based reformulation. However, other numerical algorithms can be used to solve the bilevel formulation. For instance, the bilevel program can be reformulated as a mixed-integer quadratic program (MIQP) in some cases (José Fortuny-Amat 1981, Audet et al. 1997). Our statistical results apply to any numerical algorithm for solving the bilevel formulation, including the MIQP reformulation (when possible) or our two algorithms.

We also prove that existing heuristics for inverse optimization with noisy data (Keshavarz et al. 2011, Bertsimas et al. 2014), which are expressed as convex optimization problems, are statistically inconsistent – meaning that in the limit of increasing amount of data these approaches will generate parameter estimates that converge to incorrect values. This is perhaps not unexpected, because we also prove that the problem of inverse optimization with noisy data is NP-hard. It should be noted that the inverse optimization problem *without* noisy data can be solved in polynomial time, as shown by Keshavarz et al. (2011) and Bertsimas et al. (2014).

An additional contribution is we propose a novel reformulation of bilevel programs where there lower level optimization problem is convex. It is common to replace the lower level problem by the KKT conditions or to upper bound the objective function by the value function (Dempe et al. 2015). However, these approaches face certain numerical difficulties. We propose to upper bound

the objective function by its dual, which enforces a zero duality gap and describes an optimal point. The benefit of our optimality condition is it has convexity and continuity properties that support the design of numerical algorithms. The two numerical algorithms we propose directly make use of this optimality condition, and the proofs of our statistical results are also aided by the use of this optimality condition.

1.4. Notation

Most notation we use in this paper is standard, and we briefly summarize some of the less usual aspects of our notation. We use $\|\cdot\|$ to denote the usual ℓ_2 -norm. The indicator function $\mathbb{1}(u)$ is defined to be

$$\mathbb{1}(u) = \begin{cases} 1, & \text{if condition } u \text{ is satisfied} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

When u is a vector and $\mathcal{A} = \{a_1, a_2, \dots\}$ is a set, the notation $\langle u \rangle_{\mathcal{A}}$ refers to the vector formed by the components of u with indices given in \mathcal{A} . The notation $[r] = \{1, \dots, r\}$ refers to sequential set. The Kuratowski limit superior of a sequence of sets $\mathcal{C}_\nu \subseteq \mathbb{R}^d$ is defined as

$$\limsup_\nu(\mathcal{C}_\nu) = \{x \in \mathbb{R}^d : \liminf \text{dist}(x, \mathcal{C}_\nu) = 0\}, \quad (2)$$

where $\text{dist}(x, \mathcal{C}) = \inf\{\|x - c\| \mid c \in \mathcal{C}\}$. We similarly define $\text{dist}(\mathcal{B}, \mathcal{C}) = \inf\{\text{dist}(x, \mathcal{C}) \mid x \in \mathcal{B}\}$.

2. Challenges with Noisy Inverse Optimization

This section begins by formalizing the notation for the forward problem, before defining the noisy inverse optimization problem. For the case where we have access to measurements (rather than the underlying distributions), we formulate a related sample average approximation of the inverse optimization problem. We show that both these inverse problems are NP-hard. We conclude by showing that existing heuristic approaches for solving the inverse optimization problem are statistically inconsistent, meaning that in the limit of infinite data these heuristic approaches converge to incorrect solutions.

2.1. Model for Forward Problem

Let $x \in \mathbb{R}^d$ be the decision variable, $u \in \mathbb{R}^m$ be the external input variable, and $\theta \in \mathbb{R}^p$ be the parameter vector. Then the forward optimization problem is given by

$$\text{FOP} \quad \min_x \{f(x, u, \theta) \mid g(x, u, \theta) \leq 0\},$$

where $f : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}$ is a function and $g : \mathbb{R}^d \times \mathbb{R}^m \times \mathbb{R}^p \rightarrow \mathbb{R}^q$ is a vector-valued function. The solution set of FOP is the set-valued function given by $\mathcal{S}(u, \theta) = \arg \min_x \{f(x, u, \theta) \mid g(x, u, \theta) \leq 0\}$. The value function of FOP is given by $V(u, \theta) = \min_x \{f(x, u, \theta) \mid g(x, u, \theta) \leq 0\}$, and the feasible set is defined as $\Phi(u, \theta) = \{x \in \mathbb{R}^d : g(x, u, \theta) \leq 0\}$.

2.2. Model for Inverse Optimization Problem

Suppose $(u, y) \in \mathbb{R}^m \times \mathbb{R}^d$ is a vector-valued random variable that is distributed according to some unknown but fixed joint distribution $\mathbb{P}_{(u,y)}$. Let $\mathcal{U} \times \mathcal{Y} \subseteq \mathbb{R}^m$ be the support of this distribution, meaning the smallest set that satisfies the property $\mathbb{P}_{(u,y)}(\mathcal{U}, \mathcal{Y}) = 1$. If we define the function

$$\text{RISK} \quad Q(\theta) = \mathbb{E} \left(\min_{x \in \mathcal{S}(u, \theta)} \|y - x\|^2 \right),$$

then the inverse optimization problem is given by

$$\text{IOP} \quad \min \{Q(\theta) \mid \theta \in \Theta\},$$

where $\Theta \subseteq \mathbb{R}^p$ is a known set. We make the following assumptions:

A1. The functions $f(x, u, \theta), g(x, u, \theta)$ are continuous and convex in x , for fixed u, θ .

A2. The set Θ is convex.

These assumptions are fairly mild. **A1** is equivalent to stating FOP is a convex optimization problem, and **A2** is asserting that the set of possible θ is convex.

When the joint distribution $\mathbb{P}_{(u,y)}$ is unknown, we cannot solve IOP without additional information. Fortunately, we can leverage the iid measurements (u_i, y_i) for $i \in [n]$. In the context of a decision-making agent, we should interpret the (i) u_i as an external signal the agent responds to, and (ii) y_i as a measurement of the corresponding decision of the agent. In principle, we can solve IOP using a sample average approximation:

$$\text{IOP-SAA} \quad \min \{Q_n(\theta) \mid \theta \in \Theta\},$$

where

$$\begin{aligned} \text{RISK-SAA} \quad Q_n(\theta) &= \min_{x_i} \frac{1}{n} \sum_{i=1}^n \|y_i - x_i\|^2 \\ \text{s.t. } x_i &\in \mathcal{S}(u_i, \theta), \quad \forall i \in [n] \end{aligned}$$

2.3. NP-Hardness of Inverse Optimization Problem

Though all the functions and sets involved in FOP and IOP are convex, solving IOP is NP-hard.

THEOREM 1. *If **A1, A2** hold, then IOP is NP-hard.*

Proof. We prove this by showing a reduction from the problem of computing the best rank-1 approximation of an order 3 tensor (which is NP-hard (Hillar and Lim 2013)) to IOP. Consider any

$\psi \in \mathbb{R}^{r_1 \times r_2 \times r_3}$, where $r_1, r_2, r_3 \in \mathbb{R}_+$. This defines ψ to be an order 3 tensor. We define $\rho = r_1 + r_2 + r_3$, and suppose the parameter vector is given by $\theta = (a, b, c) \in \Theta = \mathbb{R}^\rho$, where $a \in \mathbb{R}^{r_1}$, $b \in \mathbb{R}^{r_2}$, and $c \in \mathbb{R}^{r_3}$. Also define the discrete set $\mathcal{U} = [r_1] \times [r_2] \times [r_3]$, and suppose that $u = (\alpha, \beta, \gamma)$ is uniformly distributed over \mathcal{U} . Furthermore, suppose y is a random variable given by $\psi_{\alpha, \beta, \gamma}$, which means that y is dependent on u since $u = (\alpha, \beta, \gamma)$. Then we define the following forward optimization problem

$$\mathcal{S}(u, \theta) = \arg \min_x \left(x - \langle a \rangle_\alpha \cdot \langle b \rangle_\beta \cdot \langle c \rangle_\gamma \right)^2. \quad (3)$$

This forward optimization problem is a quadratic program (QP) when (u, θ) is fixed, and so the solution set is $\mathcal{S}(u, \theta) = \langle a \rangle_\alpha \langle b \rangle_\beta \langle c \rangle_\gamma$. Note that the solution set consists of a single point. Next, observe that

$$\min_{\theta \in \Theta} Q(\theta) = \min_{\theta \in \mathbb{R}^\rho} \frac{1}{\rho} \sum_{\alpha=1}^{r_1} \sum_{\beta=1}^{r_2} \sum_{\gamma=1}^{r_3} \left(\psi_{\alpha, \beta, \gamma} - \langle a \rangle_\alpha \cdot \langle b \rangle_\beta \cdot \langle c \rangle_\gamma \right)^2, \quad (4)$$

where we have converted the expectation into a weighted sum using the fact that u is uniformly distributed over \mathcal{U} . Observe that (4) is the problem of computing the best rank-1 approximation to an order 3 tensor (Hillar and Lim 2013). \square

REMARK 1. Inapproximability results for IOP can be shown under the setting where Θ is allowed to be a discrete set (i.e, **A1** holds, but **A2** does not hold). In particular, there is a straightforward reduction from the shortest vector problem. This implies that IOP is NP-hard to approximate to within any factor up to $2^{(\log d)^{1-\epsilon}}$, for any $\epsilon \geq 0$ (Haviv and Regev 2012).

REMARK 2. Polynomial-time solvability of IOP is possible in very specific settings. For instance, if the solution set of FOP is $\mathcal{S}(u, \theta) = \arg \min_x \{x^2 - 2(\theta + u) \cdot x\} = \theta + u$, then IOP is a QP: $\min_{\theta \in \Theta} \{\mathbb{E}((y - \theta - u)^2)\}$, and its minimizer is $\theta^* = \mathbb{E}(y - u)$.

This problem IOP-SAA is a bilevel program, and bilevel programs are usually difficult to solve (Dempe et al. 2015). In fact, IOP-SAA is NP-hard to solve.

COROLLARY 1. *If **A1, A2** hold, then IOP-SAA is NP-hard.*

Proof. We show this result using the same construction used to prove Theorem 1. In particular, observe that if $\{u_1, \dots, u_n\} = \mathcal{U}$, then IOP-SAA is equivalent to IOP, which is NP-hard by Theorem 1. Finally, note that the condition $\{u_1, \dots, u_n\} = \mathcal{U}$ occurs with nonzero probability since the set \mathcal{U} is finite and since the u_i are sampled uniformly from \mathcal{U} . \square

REMARK 3. Inapproximability results for IOP-SAA can be shown under the setting where Θ is allowed to be a discrete set (i.e, **A1** holds, but **A2** does not hold). In particular, the same construction in Remark 1 can be used to shown IOP-SAA is NP-hard to approximate to within any factor up to $2^{(\log d)^{1-\epsilon}}$, for any $\epsilon \geq 0$ (Haviv and Regev 2012).

REMARK 4. Polynomial-time solvability of IOP-SAA is possible in very specific settings. For instance, the construction in Remark 2 leads to an instance of IOP-SAA that is a QP.

2.4. Statistical Inconsistency of Heuristic Approaches

We begin with two statistical definitions of consistency: risk consistency and estimation consistency. These definitions are stated in order of increasing stringency, meaning that risk consistency is necessary (in situations with sufficient continuity) for estimation consistency. The first definition relates to the best predictions possible using the given forward optimization problem.

DEFINITION 1 (RISK CONSISTENCY). An estimate $\hat{\theta}_n \in \Theta$ is risk consistent if

$$Q(\hat{\theta}_n) \xrightarrow{p} \min \{Q(\theta) \mid \theta \in \Theta\}. \quad (5)$$

We should interpret the function $Q(\theta)$ as the expected prediction error when the parameter values are θ , where the prediction is the solution set $\mathcal{S}(u, \theta)$. And so the above definition is stating that an estimator θ_n is risk consistent if the expected prediction error of the estimate θ_n converges in probability to the minimum prediction error possible when we use the forward optimization model described by FOP and constrain θ to belong to Θ . In other words, an estimator is risk consistent if it asymptotically provides the best predictions possible.

The second statistical definition relates to the situation where the forward optimization model described by FOP is correct and there is a *true* parameter. In particular, it applies to situations where the below identifiability condition is satisfied. Briefly summarized, the identifiability condition is satisfied when FOP is such that two different parameter values θ_1 and θ_2 lead to two different distributions for measurements of the decision data y_i . More details and clarifying examples are found in Appendix B.

IC. There exists a unique $\theta_0 \in \Theta$ such that the following three sub-conditions hold: (i) $y = \xi + w$, where $\xi \in \mathcal{S}(u, \theta_0)$, $\mathbb{E}(w) = 0$, $\mathbb{E}(w^2) < +\infty$, and u, ξ are independent of w , (ii) for all $\theta \in \Theta \setminus \theta_0$ there exists $\mathcal{U}(\theta) \subseteq \mathcal{U}$ such that $\mathbb{P}(u \in \mathcal{U}(\theta)) > 0$ and $\text{dist}(\mathcal{S}(u, \theta), \mathcal{S}(u, \theta_0)) > 0$ for each $u \in \mathcal{U}(\theta)$, and (iii) for each fixed $\theta \in \Theta$ we have $\mathbb{P}(\{u : \mathcal{S}(u, \theta) \text{ is multivalued}\}) = 0$.

The first sub-condition of the identifiability condition is stating that the solution data y_i is a noisy measurement (with noise random variable w) of a point that belongs to the solution set $\mathcal{S}(u_i, \theta_0)$, and the second sub-condition is stating that when θ is different from θ_0 then this leads to different solution sets. This second sub-condition is necessary, because otherwise we could not distinguish the predictions of FOP when the parameters θ differ from θ_0 . The third sub-condition eliminates pathological cases that occur when the solution set at a fixed θ is so large that it approximately encompasses all possible solutions. Note that this third sub-condition is mild, and examples where it is satisfied include when (i) FOP is strictly convex, or when (ii) FOP is a linear program with random coefficients drawn from a continuous distribution; it holds for other examples as well. The second statistical definition is related to this identifiability condition.

DEFINITION 2 (ESTIMATION CONSISTENCY). Suppose **IC** holds. An estimate $\hat{\theta}_n \in \Theta$ is estimation consistent if

$$\hat{\theta}_n \xrightarrow{p} \theta_0. \quad (6)$$

Stated in words, an estimate $\hat{\theta}_n$ is estimation consistent if it converges in probability to the true parameter values θ_0 . This is the classical notion of consistency of a statistical estimator (Bickel and Doksum 2006).

Though these statistical notions of consistency are quite natural, it is the case that existing heuristic approaches for solving the inverse optimization problem are statistically inconsistent. We will use **VIA** to refer to the variational inequality method of Bertsimas et al. (2014), and we refer to the KKT conditions approach of Keshavarz et al. (2011) as **KKA**.

PROPOSITION 1. Suppose **A1**, **A2** and **IC** hold. Then **VIA** (Bertsimas et al. 2014) and **KKA** (Keshavarz et al. 2011) are not estimation consistent.

Proof. We show this using a counterexample. Suppose **FOP** is $\min\{x^2 - (\theta + u) \cdot x \mid x \in [0, 10]\}$, and note its solution set $\mathcal{S}(u, \theta) = \min\{\frac{u+\theta}{2}, 10\}$ is single-valued. Assume the distribution of u is

$$u = \begin{cases} 0, & \text{with probability (w.p.) } \frac{1}{2} \\ 20, & \text{w.p. } \frac{1}{2} \end{cases} \quad (7)$$

and that the distribution of w is

$$w = \begin{cases} -1, & \text{w.p. } \frac{1}{2} \\ +1, & \text{w.p. } \frac{1}{2} \end{cases} \quad (8)$$

Finally, suppose $y = \mathcal{S}(u, \theta) + w$, $\Theta = \{\theta \in \mathbb{R} : 0 \leq \theta \leq 10\}$, and $\theta_0 = 10$. By construction, this problem satisfies **A1**, **A2**, **IC**. Also, observe that the joint distribution of (u, y) is

$$(u, y) = \begin{cases} (0, 4), & \text{w.p. } \frac{1}{4} \\ (0, 6), & \text{w.p. } \frac{1}{4} \\ (20, 9), & \text{w.p. } \frac{1}{4} \\ (20, 11), & \text{w.p. } \frac{1}{4} \end{cases} \quad (9)$$

We show that both **VIA** and **KKA** are not estimation consistent for this problem.

We begin with **VIA**. This approach solves

$$\begin{aligned} \min_{\theta \in \Theta} \quad & \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 \\ \text{s.t.} \quad & \nabla f(y_i, u_i, \theta) \cdot (x_i - y_i) \geq -\epsilon_i, \forall x_i \in [0, 10], \quad \forall i \in [n] \end{aligned} \quad (10)$$

The constraint

$$\nabla f(y_i, u_i, \theta) \cdot (x_i - y_i) \geq -\epsilon_i, \forall x_i \in [0, 10] \quad (11)$$

is a variational inequality, and **VIA** exactly reformulates this using linear duality. We operate with the original variational inequality since the reformulation in **VIA** is exact and does not change the

solution. If $y_i = 4$, then a straightforward calculation gives that (11) is equivalent to the constraint: $\epsilon_i \geq 4 \cdot (8 - \theta)$ if $\theta \leq 8$, and $\epsilon_i \geq -6 \cdot (8 - \theta)$ if $\theta > 8$. If $y_i = 6$, then (11) is equivalent to the constraint $\epsilon_i \geq 6 \cdot (12 - \theta)$. If $y_i = 9$, then (11) is equivalent to the constraint $\epsilon_i \geq 2 + \theta$. Finally, if $y_i = 11$, then (11) is equivalent to the constraint: $\epsilon_i \geq 11 \cdot (2 - \theta)$ if $\theta \leq 2$, and $\epsilon_i \geq 2 - \theta$ if $\theta > 2$. Next, we solve the problem $\min\{\epsilon_i^2 \mid (11)\}$ for each possible value of y_i and θ . If $y_i = 4$, then the minimum is $16 \cdot (8 - \theta)^2$ if $\theta \leq 8$, and $36 \cdot (8 - \theta)^2$ if $\theta > 8$. If $y_i = 6$, then the minimum is $36 \cdot (12 - \theta)^2$. If $y_i = 9$, then minimum is $(2 + \theta)^2$. If $y_i = 11$, then the minimum is $121 \cdot (2 - \theta)^2$ if $\theta \leq 2$, and 0 if $\theta > 2$. Thus, we have

$$4 \cdot \mathbb{E}(\epsilon_i^2) = \begin{cases} 36 \cdot (12 - \theta)^2 + (2 + \theta)^2 + 121 \cdot (2 - \theta)^2 + 16 \cdot (8 - \theta)^2, & \text{if } \theta \leq 2 \\ 36 \cdot (12 - \theta)^2 + (2 + \theta)^2 + 16 \cdot (8 - \theta)^2, & \text{if } \theta \in (2, 8] \\ 36 \cdot (12 - \theta)^2 + (2 + \theta)^2 + 36 \cdot (8 - \theta)^2, & \text{if } \theta > 8 \end{cases} \quad (12)$$

Finally, we solve the optimization problem $\min\{\mathbb{E}(\epsilon_i^2) \mid \theta \in [0, 10]\}$. A simple calculation gives that the minimum occurs at $\theta^* = \frac{718}{73} \approx 9.8356$. However, the minimizer of (10) will converge in probability to θ^* , because (i) we can exactly reformulate (10) as

$$\begin{aligned} \min_{\theta \in \Theta} \quad & \frac{1}{n} \sum_{i=1}^n \epsilon_i^2 \\ \text{s.t. } \quad & \epsilon_i^2 = \begin{cases} 16 \cdot (8 - \theta)^2 \cdot \mathbb{1}(\theta \leq 8) + 36 \cdot (8 - \theta)^2 \cdot \mathbb{1}(\theta > 8), & \text{if } y_i = 4 \\ 36 \cdot (12 - \theta)^2, & \text{if } y_i = 6 \\ (2 + \theta)^2, & \text{if } y_i = 9 \\ 121 \cdot (2 - \theta)^2 \cdot \mathbb{1}(\theta \leq 2), & \text{if } y_i = 11 \end{cases} \quad \forall i \in [n] \end{aligned} \quad (13)$$

which (ii) implies we can apply the uniform law of large numbers (Jennrich 1969) since ϵ_i^2 as defined in (13) is a continuous function, and thus (iii) we get convergence of the minimizer from a standard consistency result in statistics (see for instance Theorem 5.7 in (van der Vaart 2000) or Theorem 5.2.3 in (Bickel and Doksum 2006)). This shows VIA is not estimation consistent, since $\theta_0 = 10$.

Next, we consider KKA. This approach solves

$$\begin{aligned} \min_{\theta \in \Theta} \quad & \frac{1}{n} \sum_{i=1}^n \|\epsilon_i\|^2 \\ \text{s.t. } \quad & \nabla f(y_i, u_i, \theta) - \langle \lambda_i \rangle_1 + \langle \lambda_i \rangle_2 = \langle \epsilon_i \rangle_1 \\ & - \langle \lambda_i \rangle_1 \cdot y_i = \langle \epsilon_i \rangle_2 \\ & \langle \lambda_i \rangle_2 \cdot (y_i - 10) = \langle \epsilon_i \rangle_3 \\ & \lambda_i \geq 0 \end{aligned} \quad (14)$$

We first solve the problem (14), with $n = 1$, for each possible value of y_i and θ . If $y_i = 4$, then the minimum is $\frac{16}{17} \cdot (8 - \theta)^2$ if $\theta \leq 8$, and $\frac{36}{37} \cdot (8 - \theta)^2$ if $\theta > 8$. If $y_i = 6$, then the minimum is $\frac{36}{37} \cdot (12 - \theta)^2$. If $y_i = 9$, then the minimum is $\frac{1}{2} \cdot (2 + \theta)^2$. If $y_i = 11$, then the minimum is $\frac{121}{122} \cdot (2 - \theta)^2$ if $\theta \leq 2$, and $\frac{1}{2} \cdot (2 - \theta)^2$ if $\theta > 2$. Thus, we have

$$4 \cdot \mathbb{E}(\|\epsilon_i\|^2) = \begin{cases} \frac{36}{37} \cdot (12 - \theta)^2 + \frac{1}{2} \cdot (2 + \theta)^2 + \frac{121}{122} \cdot (2 - \theta)^2 + \frac{16}{17} \cdot (8 - \theta)^2, & \text{if } \theta \leq 2 \\ \frac{36}{37} \cdot (12 - \theta)^2 + \frac{1}{2} \cdot (2 + \theta)^2 + \frac{1}{1} \cdot (2 - \theta)^2 + \frac{16}{17} \cdot (8 - \theta)^2, & \text{if } \theta \in (2, 8] \\ \frac{36}{37} \cdot (12 - \theta)^2 + \frac{1}{2} \cdot (2 + \theta)^2 + \frac{1}{2} \cdot (2 - \theta)^2 + \frac{36}{37} \cdot (8 - \theta)^2, & \text{if } \theta > 8 \end{cases} \quad (15)$$

Finally, we solve the optimization problem $\min\{\mathbb{E}(\|\epsilon_i\|^2) \mid \theta \in [0, 10]\}$. A simple calculation gives that the minimum occurs at $\theta^* = \frac{12080}{1833} \approx 6.5903$. However, the minimizer of (14) will converge in probability to θ^* , because (i) we can exactly reformulate (14) as

$$\begin{aligned} \min_{\theta \in \Theta} \quad & \frac{1}{n} \sum_{i=1}^n \|\epsilon_i\|^2 \\ \text{s.t. } \quad & \|\epsilon_i\|^2 = \begin{cases} \frac{16}{17} \cdot (8 - \theta)^2 \cdot \mathbb{1}(\theta \leq 8) + \frac{36}{37} \cdot (8 - \theta)^2 \cdot \mathbb{1}(\theta > 8), & \text{if } y_i = 4 \\ \frac{36}{37} \cdot (12 - \theta)^2, & \text{if } y_i = 6 \\ \frac{1}{2} \cdot (2 + \theta)^2, & \text{if } y_i = 9 \\ \frac{121}{122} \cdot (2 - \theta)^2 \cdot \mathbb{1}(\theta \leq 2) + \frac{1}{2} \cdot (2 - \theta)^2, & \text{if } y_i = 11 \end{cases} \quad \forall i \in [n] \end{aligned} \quad (16)$$

which (ii) implies we can apply the uniform law of large numbers (Jennrich 1969) since $\|\epsilon_i\|^2$ as defined in (16) is a continuous function, and thus (iii) we get convergence of the minimizer from a standard consistency result in statistics (see for instance Theorem 5.7 in (van der Vaart 2000) or Theorem 5.2.3 in (Bickel and Doksum 2006)). This shows that KKA is not estimation consistent, since $\theta_0 = 10$. \square

COROLLARY 2. *Suppose **A1, A2** hold. Then VIA (Bertsimas et al. 2014) and KKA (Keshavarz et al. 2011) are not risk consistent.*

Proof. Risk consistency is necessary for estimation consistency when $Q(\theta)$ is continuous. For the counterexample in the proof of Proposition 1, we have

$$Q(\theta) = \mathbb{E}\left(\|y - \min\{\frac{u+\theta}{2}, 10\}\|^2\right) = \frac{1}{4} \cdot \left((4 - \frac{\theta}{2})^2 + (6 - \frac{\theta}{2})^2 + (9 - 10)^2 + (11 - 10)^2\right) \quad (17)$$

since $\Theta = \{\theta \in \mathbb{R} : 0 \leq \theta \leq 10\}$. This $Q(\theta)$ is continuous, and so the corollary follows from Proposition 1. (As an aside, note that $\arg \min\{Q(\theta) \mid \theta \in \Theta\} = 10$, which is the correct parameter value.) \square

The intuition for why VIA and KKA are statistically inconsistent is that they are minimizing an incorrect measure of error: These approaches generate an estimated set of parameters that minimizes the level of suboptimality of the measured solution data. However, this leads to biased estimates because suboptimality is measured by (i) deviations in the value of the objective function of FOP and (ii) the amount of constraint violation of FOP, whereas noise directly perturbs the solution data. This is in contrast to our approach (as exemplified by IOP-SAA) which generate an estimated set of parameters that minimizes the deviation between predicted and measured solution data. This distinction between suboptimality and deviations in the solution data becomes most apparent (and critical) in problems with constraints.

3. Consistent Estimation for Inverse Optimization Problem

Given the statistical inconsistency of existing heuristics, we propose to solve the noisy inverse optimization problem by instead solving SAA-IOP. First, we will need to impose a regularity

condition to ensure that FOP and IOP–SAA are numerically well-posed:

R1. For each $u \in \mathcal{U}$ and $\theta \in \Theta$, the feasible set $\Phi(u, \theta)$ is closed, bounded, and has a nonempty relative interior (i.e., $\text{relint}(\Phi(u, \theta)) \neq \emptyset$). The feasible set $\Phi(u, \theta)$ is also absolutely bounded, meaning there exists $M > 0$ such that $\|x\| \leq M$, for all $x \in \Phi(u, \theta)$, $u \in \mathcal{U}$, and $\theta \in \Theta$.

Condition **R1** is equivalent to requiring FOP has a strictly feasible point (i.e., Slater’s condition holds), and that the feasible set of FOP is closed and bounded. The first sub-condition requiring the feasible set be closed and bounded is needed to ensure the existence of well-posed primal and dual solutions, and it could be replaced by more general conditions. For instance, we could have instead assumed FOP satisfies the uniform level-boundedness condition (Rockafellar and Wets 1998). We use the above for simplicity of stating the results. The second sub-condition is needed to ensure we can apply the Berge Maximum Theorem (Berge 1963).

The simplest case of statistical consistency of SAA-IOP occurs when the function $f(x, u, \theta)$ is strictly convex, because of the following result:

PROPOSITION 2. *Suppose **A1**, **A2** and **R1** hold. If $f(x, u, \theta)$ is strictly convex in x for fixed $u \in \mathcal{U}$ and $\theta \in \Theta$, then $Q_n(\theta)$ is continuous.*

Proof. Because the feasible set $\Phi(u, \theta)$ is convex for fixed u, θ by **A1** and has a nonempty interior by **R1**, this means $\Phi(u, \theta)$ is continuous in θ by Example 5.10 from (Rockafellar and Wets 1998). Thus, we can apply the Berge Maximum Theorem (Berge 1963) to FOP. This implies $\mathcal{S}(u, \theta)$ is upper hemicontinuous in θ for fixed $u \in \mathcal{U}$. However, $\mathcal{S}(u, \theta)$ consists of a single point for fixed $u \in \mathcal{U}$ and $\theta \in \Theta$, because the objective function is strictly convex and since **R1** holds. Consequently, $\mathcal{S}(u, \theta)$ is a continuous single-valued function for fixed $u \in \Theta$ (see for instance Theorem 2.6 in (Rockafellar and Wets 1998)). Thus, we can apply the Berge Maximum Theorem to RISK–SAA, and this implies that $Q_n(\theta)$ as defined in RISK–SAA is continuous. \square

In this case, we can prove risk and estimation consistency using standard arguments (Jennrich 1969, van der Vaart 2000, Bickel and Doksum 2006) from statistics that use the uniform law of large numbers (Jennrich 1969). However, this approach cannot be applied to the more general case where $f(x, u, \theta)$ is not strictly convex. In particular, when $f(x, u, \theta)$ is not strictly convex, the function $Q_n(\theta)$ will not generally be continuous. And so a different argument is required because the uniform law of large numbers does not apply to discontinuous functions.

Our approach will be to use a statistical consistency result originally due to Wald (1949) that uses a one-sided bounding argument. The advantage of this approach is that it only requires lower semicontinuity, which we show always holds for $Q_n(\theta)$. However, this result only implies

the estimates $\hat{\theta}_n$ converge in probability to the set of minimizers of $Q(\theta)$. This cannot imply risk consistency in the general case because $Q_n(\theta)$ is lower semicontinuous, which means that $Q(\hat{\theta}_n)$ can remain bounded from the minimum $Q(\theta)$. And so for the general case, we will show that a weak risk consistency result holds.

To develop the statistical consistency results for the most general case, we will develop a regularized version of RISK-SAA that is guaranteed to be continuous. The first step of this construction involves proposing a new reformulation for bilevel programs that we call a duality-based reformulation. Next, we use this reformulation to construct a regularized version of RISK-SAA and prove its continuity. We use this regularized version to prove statistical consistency results about IOP-SAA and a regularized version of IOP-SAA.

3.1. Duality-Based Reformulation

One approach to solving bilevel problems (such as IOP-SAA) is to reformulate the problem as a normal (i.e., single level) optimization problem by replacing the constraints $x_i \in \mathcal{S}(u_i, \theta)$ with an optimality condition (Dempe et al. 2015). One possibility is to replace $x_i \in \mathcal{S}(u_i, \theta)$ by the KKT conditions of FOP, and another possibility is to upper bound the objective function using the value function $f(x_i, u_i, \theta) \leq V(u_i, \theta)$. Unfortunately, these approaches often encounter numerical difficulties. The KKT approach leads to a nonlinear program with combinatorial complexity, because of the complimentary slackness in KKT. The value function approach is difficult to implement because closed-form expressions for the value function are not available except for very special cases.

Here, we present a new optimality condition. Given the numerical difficulties of existing approaches, we propose to solve bilevel programs (such as IOP-SAA) by using the Lagrangian dual function to upper bound the objective function. The following proposition shows that our idea of using the dual as an upper bound represents a novel optimality condition.

PROPOSITION 3. *Suppose **A1**, **A2** and **R1** hold. Then a point is optimal $x \in \mathcal{S}(u, \theta)$ if and only if there exists a corresponding $\lambda \in \mathbb{R}^q$ for which x, λ satisfy the inequalities*

$$\begin{aligned} f(x, u, \theta) - h(\lambda, u, \theta) &\leq 0 \\ g(x, u, \theta) &\leq 0 \\ \lambda &\geq 0 \end{aligned} \tag{18}$$

where $h(\lambda, u, \theta)$ is the Lagrangian dual function of FOP.

Proof. We first prove the forward direction. Consider any optimal point $x \in \mathcal{S}(u, \theta)$, and note that it satisfies $g(x, u, \theta) \leq 0$ since the feasible set is nonempty by **R1**. Conditions **A1**, **R1** imply strong duality, meaning there exists $\lambda \geq 0$ such that $h(\lambda, u, \theta) = f(x, u, \theta)$ (see for instance Theorem 2.165 in (Bonnans and Shapiro 2000)). As a result, this x, λ satisfies (18).

Next, we prove the reverse direction of the result. Let x be a point that satisfies (18), and note that for fixed u, θ we have

$$f(x, u, \theta) \leq h(\lambda, u, \theta) \leq \max_{\lambda} \{h(\lambda, u, \theta) \mid \lambda \geq 0\} \leq \min_x \{f(x, u, \theta) \mid g(x, u, \theta) \leq 0\}, \quad (19)$$

where the first inequality is a restatement of (18), and the last inequality follows by weak duality (e.g., (2.268) in (Bonnans and Shapiro 2000)). Hence, $x \in \mathcal{S}(u, \theta)$ (i.e., x is an optimal point). \square

As a result, we can exactly reformulate RISK-SAA as the following optimization problem:

$$\begin{aligned} Q_n(\theta) &= \min_{x_i, \lambda_i} \frac{1}{n} \sum_{i=1}^n \|y_i - x_i\|^2 \\ \text{DB-RISK-SAA} \quad &\text{s.t. } f(x_i, u_i, \theta) - h(\lambda_i, u_i, \theta) \leq 0, \quad \forall i \in [n] \\ &g(x_i, u_i, \theta) \leq 0, \quad \forall i \in [n] \\ &\lambda_i \geq 0, \quad \forall i \in [n] \end{aligned}$$

One important feature of this reformulation is that it is a convex optimization problem for fixed values of θ .

PROPOSITION 4. *Suppose **A1, R1** hold. Then DB-RISK-SAA is a convex optimization problem for fixed θ .*

Proof. Recall $f(x_i, u_i, \theta), g(x_i, u_i, \theta)$ are convex in x_i for fixed u_i, θ , by **A1**. Furthermore, the Lagrangian dual function $h(\lambda_i, u_i, \theta)$ is concave in λ_i for fixed u_i, θ (see for instance Proposition 11.48 in (Rockafellar and Wets 1998)). Moreover, there exists $\lambda_i \geq 0$ such that $h(\lambda_i, u_i, \theta)$ is finite, because **A1, R1** holds (see for instance Theorem 2.165 in (Bonnans and Shapiro 2000)). Thus, $f(x_i, u_i, \theta) - h(\lambda_i, u_i, \theta)$ is a convex function in x_i for fixed u_i, θ . Since the objective function is convex, this means that DB-RISK-SAA is a convex optimization problem. \square

3.2. Regularized Formulation

Recall that $Q_n(\cdot)$ is generally not continuous even when **A1, A2, R1** hold. Consequently, we develop a regularized version of the duality-based problem that is guaranteed to be continuous. We define the ϵ -regularized version of the duality-based problem to be

$$\begin{aligned} Q_n(\theta; \epsilon) &= \min_{x_i, \lambda_i} \frac{1}{n} \sum_{i=1}^n \|y_i - x_i\|^2 \\ \text{R-DB-RISK-SAA} \quad &\text{s.t. } f(x_i, u_i, \theta) - h(\lambda_i, u_i, \theta) \leq \epsilon, \quad \forall i \in [n] \\ &g(x_i, u_i, \theta) \leq \epsilon, \quad \forall i \in [n] \\ &\lambda_i \geq 0, \quad \forall i \in [n] \end{aligned}$$

And we associate this to a regularized version of the sample average approximation of the inverse optimization problem:

$$\text{R-IOP-SAA} \quad \min\{Q_n(\theta; \epsilon) \mid \theta \in \Theta\}.$$

The idea of this regularization is that we relax the optimality conditions to allow points x_i to be an ϵ -solution. Recall that a point

$$x^\epsilon \in \epsilon\text{-arg min}\{f(x) \mid g(x) \leq 0\}, \quad (20)$$

if (i) $f(x^\epsilon) - f^* \leq \epsilon$ and (ii) $g(x^\epsilon) \leq \epsilon$, where $f^* = \min\{f(x) \mid g(x) \leq 0\}$.

PROPOSITION 5. *Suppose **A1**, **A2** and **R1** hold. Then a point x is an ϵ -solution if and only if there exists a corresponding $\lambda \in \mathbb{R}^q$ for which x, λ satisfy the inequalities*

$$\begin{aligned} f(x, u, \theta) - h(\lambda, u, \theta) &\leq \epsilon \\ g(x, u, \theta) &\leq \epsilon \\ \lambda &\geq 0 \end{aligned} \quad (21)$$

where $h(\lambda, u, \theta)$ is the Lagrangian dual function of FOP.

Proof. We first prove the forward direction. Consider any point x that is an ϵ -solution, and note it satisfies $g(x, u, \theta) \leq \epsilon$ by definition. Conditions **A1**, **R1** imply strong duality, meaning there exists $\lambda \geq 0$ such that $h(\lambda, u, \theta) = \min\{f(x, u, \theta) \mid g(x, u, \theta) \leq 0\}$ (see for instance Theorem 2.165 in (Bonnans and Shapiro 2000)). Combining this with the definition of an ϵ -solution, we have

$$f(x, u, \theta) - h(\lambda, u, \theta) = f(x, u, \theta) - \min\{f(x, u, \theta) \mid g(x, u, \theta) \leq 0\} \leq \epsilon. \quad (22)$$

As a result, this x, λ satisfies (18).

Next, we prove the reverse direction of the result. Let x be a point that satisfies (21), and note that for fixed u, θ we have

$$f(x, u, \theta) \leq h(\lambda, u, \theta) + \epsilon \leq \max_{\lambda} \{h(\lambda, u, \theta) \mid \lambda \geq 0\} + \epsilon \leq \min_x \{f(x, u, \theta) \mid g(x, u, \theta) \leq 0\} + \epsilon, \quad (23)$$

where the first inequality is a restatement of (21), and the last inequality follows from weak duality (e.g., (2.268) in (Bonnans and Shapiro 2000)). This means that x must be an ϵ -solution. \square

One benefit of this regularization is that it ensures convexity of R-DB-RISK-SAA when θ is fixed.

PROPOSITION 6. *Suppose **A1**, **A2** and **R1** hold. Then R-DB-RISK-SAA is a convex optimization problem for fixed θ .*

Proof. The proof is identical to that of Proposition 6. \square

Though the above propositions show that the regularization is equivalent to replacing optimality conditions with ϵ -optimality conditions while maintaining convexity for fixed values of θ , the main benefit of the regularization is that it ensures the function $Q_n(\theta; \epsilon)$ defined in R-DB-RISK-SAA is continuous in θ, ϵ for any $\epsilon > 0$.

PROPOSITION 7. *Suppose **A1**, **A2** and **R1** hold. Then the function $Q_n(\theta; \epsilon)$ is jointly continuous in θ, ϵ for any $\epsilon > 0$.*

Proof. The solution set $\mathcal{S}(u, \theta)$ is nonempty under **A1**, **R1** (see for instance Theorem 1.9 of (Rockafellar and Wets 1998)). Pick any $x_i \in \mathcal{S}(u, \theta)$, and let λ_i be such that x_i, λ_i satisfy (18) – this λ_i exists by Proposition 3. Next, consider the sets

$$\begin{aligned}\bar{\mathcal{S}}(u_i, \theta; \epsilon) &= \{x : f(x, u_i, \theta) - h(\lambda_i, u_i, \theta) \leq \epsilon, g(x, u_i, \theta) \leq \epsilon\} \\ \mathcal{S}(u_i, \theta; \epsilon) &= \{x : f(x, u_i, \theta) - h(\lambda, u_i, \theta) \leq \epsilon, g(x, u_i, \theta) \leq \epsilon, \lambda \geq 0\}\end{aligned}\tag{24}$$

and note that $\bar{\mathcal{S}}(u_i, \theta; \epsilon) = \mathcal{S}(u_i, \theta; \epsilon)$, since as shown in the proof of Proposition 3 we have $h(\lambda, u_i, \theta) \leq h(\lambda_i, u_i, \theta)$ for all $\lambda \geq 0$. Observe that the functions $f(x_i, u_i, \theta), g(x_i, u_i, \theta)$ are continuous and convex from **A1**, and the point x_i belongs to the interior of $\bar{\mathcal{S}}(u_i, \theta; \epsilon)$ since it satisfies (18). Thus, we can apply Example 5.10 from (Rockafellar and Wets 1998): This yields that $\bar{\mathcal{S}}(u_i, \theta; \epsilon)$ is continuous in θ, ϵ for any $\epsilon > 0$, and so we also get continuity of $\mathcal{S}(u_i, \theta; \epsilon)$ by its equality to $\bar{\mathcal{S}}(u_i, \theta; \epsilon)$. Since R-DB-RISK-SAA can be written as $Q_n(\theta; \epsilon) = \min_{x_i} \{\frac{1}{n} \sum_{i=1}^n \|y_i - x_i\|^2 \mid x_i \in \mathcal{S}(u_i, \theta; \epsilon), \forall i \in [n]\}$, we are able to apply the Berge Maximum Theorem (Berge 1963). This implies continuity of $Q_n(\theta; \epsilon)$ in θ, ϵ for any $\epsilon > 0$. \square

A point of note is that within the above proof, we show that the set of ϵ -optimal solutions of a parametric convex optimization problem $\mathcal{S}(u_i, \theta; \epsilon)$ is continuous with respect to the parametrization θ ; this is in contrast to the solution set of a parametric convex optimization problem $\mathcal{S}(u_i, \theta)$, which is in general only upper hemicontinuous with respect to the parametrization θ . The case of a parametric strictly convex optimization problem is the exception, which as shown in the proof of Proposition 2 has a continuous (with respect to the parametrization θ) solution set.

The function $Q_n(\theta; \epsilon)$ will not be jointly continuous in θ, ϵ at $\epsilon = 0$. However, it satisfies another property that is useful for solving IOP-SAA:

PROPOSITION 8. *Suppose **A1**, **A2** and **R1** hold, and let $\epsilon_\nu > 0$ be a monotone decreasing sequence with $\epsilon_\nu \rightarrow 0$. Then we have $\min\{Q_n(\theta; \epsilon_\nu) \mid \theta \in \Theta\} \rightarrow \min\{Q_n(\theta) \mid \theta \in \Theta\}$ and*

$$\limsup_\nu (\arg \min\{Q_n(\theta; \epsilon_\nu) \mid \theta \in \Theta\}) \subseteq \arg \min\{Q_n(\theta) \mid \theta \in \Theta\}.\tag{25}$$

If $z_\nu > 0$ is a monotone decreasing sequence with $z_\nu \rightarrow 0$, then we also have

$$\limsup_\nu (z_\nu - \arg \min\{Q_n(\theta; \epsilon_\nu) \mid \theta \in \Theta\}) \subseteq \arg \min\{Q_n(\theta) \mid \theta \in \Theta\}.\tag{26}$$

Proof. Let $\mathcal{C}_n(\theta, \epsilon)$ be the feasible set of R-DB-RISK-SAA, and define $(X, \Lambda) = \{x_i, \lambda_i, \forall i \in [n]\}$. Suppose $(X, \Lambda) \in \mathcal{C}_n(\theta, \alpha)$, where $\alpha \geq 0$. Then for any $\beta \geq \alpha$ we must have $(X, \Lambda) \in \mathcal{C}_n(\theta, \beta)$ by the definition of the constraints in R-DB-RISK-SAA. This means that

$$\mathcal{C}_n(\theta, \epsilon_1) \supseteq \mathcal{C}_n(\theta, \epsilon_2) \supseteq \dots\tag{27}$$

As a result, the set $\mathcal{D}_n(\theta, \epsilon_\nu) = \{\theta, X, \Lambda : \theta \in \Theta \text{ and } (X, \Lambda) \in \mathcal{C}_n(\theta, \epsilon_\nu)\}$ is also monotone nonincreasing:

$$\mathcal{D}_n(\theta, \epsilon_1) \supseteq \mathcal{D}_n(\theta, \epsilon_2) \supseteq \cdots \quad (28)$$

Also, the feasible set $\Phi(u, \theta)$ is convex for fixed u, θ by **A1** and has a nonempty interior by **R1**. This means $\Phi(u, \theta)$ is continuous in θ by Example 5.10 from (Rockafellar and Wets 1998), and so we can apply the Berge Maximum Theorem (Berge 1963) to FOP. This implies $\mathcal{S}(u, \theta)$ is upper hemicontinuous in θ for fixed $u \in \mathcal{U}$. By Remark 3.2 of (Dempe et al. 2015), this means $Q_n(\theta)$ is lower semicontinuous. Thus, by Proposition 7.4.d of (Rockafellar and Wets 1998) we have that the extended real-valued function $\{Q_n(\theta; \epsilon_\nu) \mid \theta \in \Theta\}$ epiconverges to the extended real-valued function $\{Q_n(\theta) \mid \theta \in \Theta\}$. The result then follows from Exercise 7.32.d and Theorem 7.33 of (Rockafellar and Wets 1998). \square

COROLLARY 3. *Suppose **A1**, **A2** and **R1** hold. Given any $d > 0$, there exists $E, Z > 0$ such that if $\hat{\theta}_n \in z\text{-arg min}\{Q_n(\theta; \epsilon) \mid \theta \in \Theta\}$ for any $0 \leq z \leq Z$ and $0 \leq \epsilon \leq E$, then $\text{dist}(\hat{\theta}_n, \arg \min\{Q_n(\theta) \mid \theta \in \Theta\}) < d$.*

Proof. This is a restatement of Proposition 8. \square

These results say that approximately solving R-IOP-SAA is equivalent to approximately solving IOP-SAA.

3.3. Statistical Consistency

In order to prove statistical consistency, we will need to impose an additional regularity condition that ensures expectations of corresponding random variables exist.

R2. The set Θ is closed and bounded, and $\mathbb{E}(y^2) < +\infty$.

This regularity assumption ensures that the law of large numbers (Wald 1949, Jennrich 1969, van der Vaart 2000) holds in our setting. The above expectation condition holds in many situations, including when \mathcal{Y} is bounded or when y has a sub-exponential distribution (Vershynin 2012). This allows for settings where **IC** holds with measurement noise that is Gaussian, Bernoulli, bounded support, Laplacian, Exponential, and many other distributions.

Our first statistical consistency result is that solving R-IOP-SAA is risk consistent. To state the result, we must formally define the regularized version of the inverse optimization problem. The regularized risk is

$$\text{R-RISK} \quad Q(\theta; \epsilon) = \mathbb{E}\left(\min_{x \in \mathcal{S}(u, \theta; \epsilon)} \|y - x\|^2\right),$$

where $\mathcal{S}(u, \theta; \epsilon) = \{x \in \mathbb{R}^d : f(x, u, \theta) \leq V(u, \theta) + \epsilon, g(x, u, \theta) \leq \epsilon\}$ is the set of ϵ -solutions to FOP. We define the regularized inverse optimization problem to be

$$\text{R-IOP} \quad \min\{Q(\theta; \epsilon) \mid \theta \in \Theta\}.$$

The first statistical consistency result specifically concerns nearly-optimal solutions of R-IOP-SAA. We say that a sequence of solutions $\hat{\theta}_n$ is nearly-optimal for R-IOP-SAA with fixed $\epsilon > 0$ in probability if for any $\delta > 0$ we have

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\text{dist}(\hat{\theta}_n, \arg \min\{Q_n(\theta; \epsilon) \mid \theta \in \Theta\}) > \delta\right) = 0. \quad (29)$$

THEOREM 2. *Suppose **A1**, **A2** and **R1**, **R2** hold. Given any fixed $\epsilon > 0$, if $\hat{\theta}_n$ is nearly-optimal for R-IOP-SAA with high probability, then we have $Q(\hat{\theta}_n; \epsilon) \xrightarrow{p} \min\{Q(\theta; \epsilon) \mid \theta \in \Theta\}$.*

Proof. Proposition 7 gives continuity of $Q_n(\theta; \epsilon)$. Thus, we can apply the uniform law of large numbers (Jennrich 1969), which gives

$$\sup_{\theta \in \Theta} |Q_n(\theta; \epsilon) - Q(\theta; \epsilon)| \xrightarrow{p} 0. \quad (30)$$

Consider any $\theta_0 \in \arg \min\{Q(\theta; \epsilon) \mid \theta \in \Theta\}$ and any $\theta_1 \in \arg \min\{Q_n(\theta; \epsilon) \mid \theta \in \Theta\}$. By assumption $Q_n(\theta_1; \epsilon) \leq Q_n(\theta_0; \epsilon)$, and so we have

$$Q(\hat{\theta}_n; \epsilon) + Q_n(\hat{\theta}_n; \epsilon) - Q(\hat{\theta}_n; \epsilon) + Q_n(\theta_1; \epsilon) - Q_n(\hat{\theta}_n; \epsilon) \leq Q(\theta_0; \epsilon) + Q_n(\theta_0; \epsilon) - Q(\theta_0; \epsilon). \quad (31)$$

Rearranging terms gives

$$Q(\hat{\theta}_n; \epsilon) - Q(\theta_0; \epsilon) \leq |Q_n(\hat{\theta}_n; \epsilon) - Q(\hat{\theta}_n; \epsilon)| + |Q_n(\theta_1; \epsilon) - Q_n(\hat{\theta}_n; \epsilon)| + |Q_n(\theta_0; \epsilon) - Q(\theta_0; \epsilon)|. \quad (32)$$

Recall (i) $Q(\theta_0; \epsilon) \leq Q(\hat{\theta}_n; \epsilon)$ by definition of θ_0 , (ii) $Q_n(\theta; \epsilon)$ is continuous, and (iii) $\hat{\theta}_n$ is nearly-optimal for R-IOP-SAA with high probability. Thus, combining these facts with (30) and (32) gives that $Q(\hat{\theta}_n; \epsilon) - Q(\theta_0; \epsilon) \xrightarrow{p} 0$. This is the desired result. \square

This result says that if choose any $\epsilon > 0$ and solve R-IOP-SAA to generate an estimate $\hat{\theta}_n$, then the predictions given by the ϵ -solutions to FOP (i.e., $\mathcal{S}(u, \hat{\theta}_n; \epsilon)$) are asymptotically the best possible set of predictions when the error of predictions is measured using R-RISK. A stronger risk consistency result is not possible in the general setting because $Q(\theta)$ is typically discontinuous, and so the above result can be interpreted as a weak consistency result.

A stronger risk consistency result is possible in the case where $f(x, u, \theta)$ is strictly convex. We say that a sequence of solutions $\hat{\theta}_n$ is nearly-optimal for IOP-SAA in probability if for any $\delta > 0$ we have

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(\text{dist}(\hat{\theta}_n, \arg \min\{Q_n(\theta) \mid \theta \in \Theta\}) > \delta\right) = 0. \quad (33)$$

THEOREM 3. Suppose **A1,A2** and **R1,R2** hold. If $f(x, u, \theta)$ is strictly convex in x (for fixed $u \in \mathcal{U}$ and $\theta \in \Theta$) and $\hat{\theta}_n$ is nearly-optimal for IOP-SAA with high probability, then we have $Q(\hat{\theta}_n) \xrightarrow{p} \min \{Q(\theta) \mid \theta \in \Theta\}$.

Proof. Proposition 2 gives continuity of $Q_n(\theta)$. The remainder of the proof is identical to Theorem 2. \square

This result says that when FOP is a strictly convex optimization problem and we solve IOP-SAA to generate an estimate $\hat{\theta}_n$, then the predictions given by the solutions to FOP (i.e., $\mathcal{S}(u, \hat{\theta}_n)$) are asymptotically the best possible set of predictions when the error of predictions is measured using RISK. The reason it is possible to show risk consistency in this case is that $Q(\theta)$ will be continuous in this setting.

Our final statistical consistency result is that solving IOP-SAA is estimation consistent when **IC** holds.

THEOREM 4. Suppose **A1,A2** and **R1,R2** and **IC** hold. If $\hat{\theta}_n$ is nearly-optimal for IOP-SAA with high probability, then we have $\hat{\theta}_n \xrightarrow{p} \theta_0$.

Proof. Because the feasible set $\Phi(u, \theta)$ is convex for fixed u, θ by **A1** and has a nonempty interior by **R1**, this means $\Phi(u, \theta)$ is continuous in θ by Example 5.10 from (Rockafellar and Wets 1998). Thus, we can apply the Berge Maximum Theorem (Berge 1963) to FOP. This implies $\mathcal{S}(u, \theta)$ is upper hemicontinuous in θ for fixed $u \in \mathcal{U}$. By Remark 3.2 of Dempe et al. (2015), this means $Q_n(\theta)$ is lower semicontinuous. Thus, we can apply Theorem 5.14 of (van der Vaart 2000). (Technically, this theorem applies to maximizing upper semicontinuous functions, but the results and proof trivially extend to the case of minimizing lower semicontinuous functions.) The result follows from the conclusion of Theorem 5.14 of (van der Vaart 2000) if we can show (i) $\theta_0 \in \arg \min \{Q(\theta) \mid \theta \in \Theta\}$, and that (ii) θ_0 is the unique solution. First, note $Q(\theta) = \mathbb{E}(\min_{x \in \mathcal{S}(u, \theta)} \|\xi - x\|^2) + \mathbb{E}(w^2)$, since ξ, x is almost surely independent of w because by **IC** we have that (i) ξ, u are independent of w , and (ii) $\mathcal{S}(u, \theta)$ is almost surely single-valued. Since by **IC** we have $\xi \in \mathcal{S}(u, \theta_0)$, this means that $Q(\theta_0) = \mathbb{E}(w^2)$ and that $\theta_0 \in \arg \min \{Q(\theta) \mid \theta \in \Theta\}$. Next, consider any $\theta \in \Theta \setminus \theta_0$. Then by **IC** we have $\mathbb{E}[\min_{x \in \mathcal{S}(u, \theta)} \|\xi - x\|^2 \mid u \in \mathcal{U}(\theta)] > 0$ since $\xi \in \mathcal{S}(u, \theta_0)$ and $\text{dist}(\mathcal{S}(u, \theta), \mathcal{S}(u, \theta_0)) > 0$ for each $u \in \mathcal{U}(\theta)$. Because $\mathbb{P}(u \in \mathcal{U}(\theta)) > 0$ from **IC**, this means $\mathbb{E}(\min_{x \in \mathcal{S}(u, \theta)} \|\xi - x\|^2) > 0$ for any $\theta \in \Theta \setminus \theta_0$. Consequently, we have $Q(\theta) > Q(\theta_0)$ for any $\theta \in \Theta \setminus \theta_0$. \square

4. Numerical Approaches to Solving IOP-SAA

Solving IOP-SAA with $Q_n(\theta)$ as formulated in DB-RISK-SAA is still difficult because it is a non-convex problem even under **A1,A2,R1**. We will propose two approaches to solving this problem. The first is an enumeration algorithm that is applicable to situations where p is modest (i.e., the

$\theta \in \mathbb{R}^p$ parameter has between 1 to 5 dimensions). The second approach we describe is a semi-parametric algorithm, and it can be used in cases where $\theta \in \mathbb{R}^p$ is higher-dimensional and y has a specific distribution. For both algorithms, we will prove that the estimates computed by these methods satisfy the conditions required for statistical consistency.

The difference in the two algorithms is how they trade-off computational and statistical performance. The enumeration algorithm requires exponential in p computation, while the semiparametric algorithm needs polynomial in p computation. But the statistical performance of the methods will be the opposite. The estimates and risk of the enumeration algorithm are anticipated to converge at faster rate than those of the semiparametric algorithm. The reason is that the semiparametric algorithm makes use of a nonparametric step (via the L2NW estimator), which is well-known to generally converge at a slower rate than a fully parametric approach. Precisely characterizing the statistical convergence rates of the two algorithms is left open for future work.

Though the enumeration algorithm needs exponential in p computation, it is still practical for many real-world problems. Many principal-agent problems (e.g. ??) use models where the parameter set is modest in dimensionality (i.e., utility functions with 2 or 3 *type* parameters). We demonstrate the practicality of the enumeration algorithm in Section 5 through an energy-related example using real data.

4.1. Enumeration Algorithm

The main idea of this algorithm is that computing $Q_n(\theta)$ and $Q_n(\theta; \epsilon)$ for fixed values of θ can be done in polynomial time since DB-RISK-SAA and R-DB-RISK-SAA are convex optimization problems by Propositions 4 and 6, respectively. This approach enumerates over different fixed values of θ and solves a series of polynomial time problems. However, Θ is a continuous set since because it is convex by **A2**. To enable enumeration, we discretize Θ using a δ -net of Θ , which we will call $\mathcal{T}(\delta)$. (Here, we define this to mean that $\mathcal{T}(\delta)$ is a finite set such that $\max_{\theta \in \Theta} \min_{t \in \mathcal{T}(\delta)} \|t - \theta\| \leq \delta$.) We then compute $Q_n(\theta; \epsilon)$ for all $\theta \in \mathcal{T}(\delta)$. And our approximate solution is finally given by $\hat{\theta}_n = \arg \min \{Q_n(\theta; \epsilon) \mid \theta \in \mathcal{T}(\delta)\}$.

This approach requires continuity of $Q_n(\theta; \epsilon)$ because otherwise performing an enumeration via the δ -net $\mathcal{T}(\delta)$ may not get sufficiently close to the optimal value. However, $Q_n(\theta; \epsilon)$ is only guaranteed to be continuous at $\epsilon = 0$ when $f(x, u, \theta)$ is strictly convex for fixed u, θ by Proposition 2 and since $Q_n(\theta; 0) = Q_n(\theta)$ by definition. Hence, we require $\epsilon > 0$ for cases where $f(x, u, \theta)$ is *not* strictly convex to ensure continuity of $Q_n(\theta; \epsilon)$ by Proposition 7. Of course, when $f(x, u, \theta)$ is strictly convex we can set $\epsilon = 0$ and maintain continuity of $Q_n(\theta; \epsilon)$.

This approach is formally presented in Algorithm 1. Importantly, it can be shown that this enumeration algorithm generates nearly-optimal solutions of IOP-SAA and R-IOP-SAA. This means

Algorithm 1: Enumeration Algorithm**Data:** fixed $\delta > 0$ and $\epsilon \geq 0$ **Result:** estimate $\hat{\theta}_n$

- 1 set $\mathcal{T}(\delta)$ to be δ -net of Θ ;
- 2 **foreach** $\theta \in \mathcal{T}(\delta)$ **do**
- 3 compute $Q_n(\theta; \epsilon)$ by solving R-DB-RISK-SAA;
- 4 set $\hat{\theta}_n \in \arg \min\{Q_n(\theta; \epsilon) \mid \theta \in \mathcal{T}(\delta)\}$;

the statistical consistency results in Section 3.3 apply to the solutions computed by this algorithm. In practice, ϵ is chosen to be $\epsilon = 0$ when FOP is strictly convex, and otherwise ϵ is chosen to be a small positive value that controls the desired precision of the resulting estimate.

THEOREM 5. *Suppose **A1**, **A2** and **R1** hold. Given any $d > 0$, there exists $E, \Delta > 0$ such that if $\hat{\theta}_n$ is computed using the enumeration algorithm (i.e., Algorithm 1) for any $0 < \epsilon \leq E$ and $0 < \delta \leq \Delta$, then $\text{dist}(\hat{\theta}_n, \arg \min\{Q_n(\theta) \mid \theta \in \Theta\}) < d$.*

Proof. By Corollary 3, there exists $E, Z > 0$ such that if $\hat{\theta}_n \in z\text{-arg min}\{Q_n(\theta; \epsilon) \mid \theta \in \Theta\}$ for any $0 \leq z \leq Z$ and $0 \leq \epsilon \leq E$, then $\text{dist}(\hat{\theta}_n, \arg \min\{Q_n(\theta) \mid \theta \in \Theta\}) < d$. Suppose we choose $z = Z$. Because $Q_n(\theta; \epsilon)$ is continuous in θ by Proposition 7, there exists $\Delta > 0$ such that for any $0 < \delta \leq \Delta$ we have

$$\min\{Q_n(\theta; \epsilon) - Q_n(\theta_0; \epsilon) \mid \theta \in \mathcal{T}(\delta)\} < z, \quad (34)$$

where $\theta_0 \in \arg \min\{Q_n(\theta; \epsilon) \mid \theta \in \Theta\}$. By construction, we have

$$\arg \min\{Q_n(\theta; \epsilon) \mid \theta \in \mathcal{T}(\delta)\} \subseteq z\text{-arg min}\{Q_n(\theta; \epsilon) \mid \theta \in \Theta\}. \quad (35)$$

Next, note the enumeration algorithm returns a solution $\hat{\theta}_n \in \arg \min\{Q_n(\theta; \epsilon) \mid \theta \in \mathcal{T}(\delta)\}$, which also satisfies $\hat{\theta}_n \in z\text{-arg min}\{Q_n(\theta; \epsilon) \mid \theta \in \Theta\}$. The result follows from applying the first line of the proof. \square

As mentioned above, in the special case where FOP is a strictly convex optimization problem we can simplify the algorithm by setting $\epsilon = 0$. We have a corresponding result about the correctness of the algorithm in this case.

THEOREM 6. *Suppose **A1**, **A2** and **R1** hold. If $f(x, u, \theta)$ is strictly convex in x (for fixed $u \in \mathcal{U}$ and $\theta \in \Theta$), then given any $d > 0$ there exists $\Delta > 0$ such that if $\hat{\theta}_n$ is computed using the enumeration algorithm for $\epsilon = 0$ and any $0 < \delta \leq \Delta$, then $\text{dist}(\hat{\theta}_n, \arg \min\{Q_n(\theta) \mid \theta \in \Theta\}) < d$.*

Proof. By Corollary 3, there exists $E, Z > 0$ such that if $\hat{\theta}_n \in z\text{-arg min}\{Q_n(\theta; \epsilon) \mid \theta \in \Theta\}$ for any $0 \leq z \leq Z$ and $0 \leq \epsilon \leq E$, then $\text{dist}(\hat{\theta}_n, \arg \min\{Q_n(\theta) \mid \theta \in \Theta\}) < d$. Suppose we choose $z = Z$

and $\epsilon = 0$, and note that $Q_n(\theta; 0) = Q_n(\theta)$ by their definitions. Because $Q_n(\theta)$ is continuous in θ by Proposition 2, there exists $\Delta > 0$ such that for any $0 < \delta \leq \Delta$ we have

$$\min \{Q_n(\theta) - Q_n(\theta_0) \mid \theta \in \mathcal{T}(\delta)\} < z, \quad (36)$$

where $\theta_0 \in \arg \min\{Q_n(\theta) \mid \theta \in \Theta\}$. By construction, we have

$$\arg \min\{Q_n(\theta) \mid \theta \in \mathcal{T}(\delta)\} \subseteq z\text{-}\arg \min\{Q_n(\theta) \mid \theta \in \Theta\}. \quad (37)$$

Next, note the enumeration algorithm returns a solution $\hat{\theta}_n \in \arg \min\{Q_n(\theta) \mid \theta \in \mathcal{T}(\delta)\}$, which also satisfies $\hat{\theta}_n \in z\text{-}\arg \min\{Q_n(\theta) \mid \theta \in \Theta\}$. The result follows from the first line of the proof. \square

4.2. Semiparametric Approach

Our second approach to solving IOP-SAA is a semiparametric approach. We will need to make an additional assumption about the structure of the problem, as well as impose two more regularity conditions, in order to be able use this approach. We begin with the additional assumption.

A3. The constraint function $g(x, u, \theta)$ is independent of θ , meaning it can be written as $g(x, u, \theta) = g_0(x, u)$. The objective function $f(x, u, \theta)$ is affine in θ , meaning it can be written as

$$f(x, u, \theta) = f_0(x, u) + \sum_{j=1}^p \langle \theta \rangle_j f_j(x, u). \quad (38)$$

Independence of the constraint g from θ is required because the semiparametric approach relies on fully knowing the feasible region of the forward problem. We note that this is not a particularly strong assumption, since in utility estimation settings one would expect the unknown parameters to appear in the objective function of the forward problem. Keshavarz et al. (2011) and Bertsimas et al. (2014) also assume that the feasible region of the forward problem is independent of the unknown parameters. The second part of **A3** ensures that the Lagrangian dual function $h(\lambda, u, \theta)$ is concave in θ . This will enable efficient computation in our semiparametric approach. Next, we describe the two additional regularity conditions. The first is

R3. The objective function $f(x, u, \theta)$ is strictly convex in x (for fixed $u \in \mathcal{U}$ and $\theta \in \Theta$) and twice continuously differentiable in x, u, θ , and the constraints $g(x, u, \theta)$ are continuously differentiable in x, u, θ .

Condition **R3** ensures smoothness in the objective function and constraints. The reason we also include a strict convexity assumption is that it acts a regularity condition: Strictly speaking, we require the second-order growth condition

$$f(x, u, \theta) \geq V(u, \theta) + c \cdot [\text{dist}(x, \mathcal{S}(u, \theta))]^2, \quad (39)$$

for some $c > 0$ and all $x \in \Phi(u, \theta)$. Unfortunately, this condition can be difficult to check even though it has been completely characterized for convex optimization problems (Bonnans and Ioffe 1995). Fortunately, strict convexity with constraint qualification implies this second-order growth condition (Bonnans 1992). Also, note that our results could be extended to the case where the problem satisfies the first-order growth condition

$$f(x, u, \theta) \geq V(u, \theta) + c \cdot \text{dist}(x, \mathcal{S}(u, \theta)), \quad (40)$$

for some $c > 0$ and all $x \in \Phi(u, \theta)$. We do not consider this extension in the present paper.

R4. The noise random variable w has a sub-exponential distribution, meaning there exists $c > 0$ such that $\mathbb{P}(|w| > t) \leq \exp(1 - t/c)$. Also, the probability density function $\mu(u)$ of u is continuously differentiable and is bounded from zero (i.e., $\min_{u \in \mathcal{U}} \mu(u) > 0$).

This regularity condition ensures the distribution of the random variables w, u are not extreme. Most commonly used heavy-tailed noise distributions are sub-exponential distributions, and so **R4** is satisfied by Gaussian, Bernoulli, bounded support, Laplacian, Exponential, and many other distributions (Vershynin 2012). Also, the regularity condition on $\mu(u)$ implies \mathcal{U} is bounded.

The idea behind the semiparametric approach is the observation that **R-DB-RISK-SAA** is convex in θ for fixed x when **A3** holds. However, because the y_i are measured with noise, we cannot simply make the substitution $x_i = y_i$. To overcome this difficulty, we first de-noise the y_i using a nonparametric estimator. Specifically, we define the ℓ_2 -regularized Nadaraya-Watson (L2NW) estimator (Aswani et al. 2013) as

$$\bar{x}_i = \frac{\gamma^{-m} \cdot \frac{1}{n} \sum_{j=1}^n y_j \cdot K\left(\frac{u_j - u_i}{\gamma}\right)}{\sigma + \gamma^{-m} \cdot \frac{1}{n} \sum_{j=1}^n K\left(\frac{u_j - u_i}{\gamma}\right)}, \quad (41)$$

where $\gamma > 0$ is the *bandwidth* parameter, $\sigma > 0$ is the ℓ_2 -regularization parameter, and $K : \mathbb{R}^m \rightarrow \mathbb{R}$ is a *kernel function* that satisfies the following properties (i) $K(u) \geq 0$, (ii) $K(u) = 0$ for $\|u\| > 1$, (iii) $K(u) = K(-u)$, and (iv) $\int K(u) du = 1$. A common example of a kernel function is the Epanechnikov kernel, which is defined as the function

$$K(u) = \begin{cases} \frac{3}{4} \cdot (1 - \|u\|^2), & \text{if } \|u\| \leq 1 \\ 0, & \text{otherwise} \end{cases} \quad (42)$$

The L2NW estimator (41) is computed in polynomial time, and it serves to de-noise the x_i in the manner described by the following proposition.

PROPOSITION 9. Suppose **A1** and **R1–R4** hold. If $\gamma = O(n^{-2/(8m+1)})$ and $\sigma = O(\gamma)$, then $\mathcal{S}(u, \theta)$ consists of a single point, and for sufficiently large n we have

$$\mathbb{P}\left(\max_{i \in [n]} \|\bar{x}_i - \mathcal{S}(u_i, \theta_0)\| > n^{-1/(18m)}\right) \leq k_1 \exp\left(-k_2 n^{1/4}\right), \quad (43)$$

where $k_1, k_2 > 0$ are constants. In particular, this implies $\max_{i \in [n]} \|\bar{x}_i - \mathcal{S}(u_i, \theta_0)\| \xrightarrow{p} 0$.

Proof. The first part follows from the strict convexity assumption in **R2**, and the third part follows directly from the second part. And so we focus on proving the second part. We will prove this using a truncation argument (see for instance (Tao 2012)).

First, note that the function $\psi(x, y) = x/y$ over the domain $(x, y) \in [-M, M] \times [\sigma, \sigma + 1]$ is Lipschitz continuous with constant $L_1 = \sqrt{(M^2 + (\sigma + 1)^2)}/\sigma^2$. Suppose we choose $M = \max_{u \in \mathcal{U}} \|\mu(u)\mathcal{S}(u, \theta_0)\| + 1$. As a result, using Lemma 1 and Lemma 2 we have

$$\begin{aligned} & \mathbb{P}\left(\left\|\bar{x}_i - \frac{\mu(u_i)\mathcal{S}(u_i, \theta_0)}{\sigma + \mu(u_i)}\right\| > t\right) \\ & \leq \mathbb{P}\left(\left|\gamma^{-m} \cdot \frac{1}{n} \sum_{j=1}^n K\left(\frac{u_j - u_i}{\gamma}\right) - \mu(u_i)\right| > t/L_1\right) + \mathbb{P}\left(\left\|\gamma^{-m} \cdot \frac{1}{n} \sum_{j=1}^n y_j \cdot K\left(\frac{u_j - u_i}{\gamma}\right)\right\| > M\right) + \\ & \quad \mathbb{P}\left(\left\|\gamma^{-m} \cdot \frac{1}{n} \sum_{j=1}^n y_j \cdot K\left(\frac{u_j - u_i}{\gamma}\right) - \mu(u_i)\mathcal{S}(u_i, \theta_0)\right\| > t/L_1\right) \\ & \leq 2 \exp\left(-2c_2 n \gamma^{2m} \cdot (t/L_1 - c_1 \cdot \gamma)^2\right) + 2 \exp\left(-2c_2 n \gamma^{2m} \cdot (1 - c_1 \cdot \gamma)^2\right) + \\ & \quad 2 \exp\left(-2c_5 n \gamma^{2m} \cdot (t/L_1 - c_3 \cdot \gamma^{1/2} - c_4 \cdot \gamma)\right), \end{aligned} \quad (44)$$

for $t > \max\{c_1 \cdot \gamma, c_3 \cdot \gamma^{1/2} + c_4 \cdot \gamma\}$. Next, observe that the function $\psi(x, y)$ over the domain

$$(x, y) \in [\min_{u \in \mathcal{U}} \mu(u)\mathcal{S}(u, \theta), \max_{u \in \mathcal{U}} \mu(u)\mathcal{S}(u, \theta)] \times [\min_{u \in \mathcal{U}} \mu(u), \max_{u \in \mathcal{U}} \mu(u)], \quad (45)$$

is Lipschitz continuous with some constant $L_2 > 0$ since (i) the denominator of ψ is bounded away from zero because of **R4**, and (ii) the numerator of ψ is bounded by **R1, R4**. Thus, we have

$$\mathbb{P}\left(\|\bar{x}_i - \mathcal{S}(u_i, \theta_0)\| > t\right) \leq \mathbb{P}\left(\left\|\bar{x}_i - \frac{\mu(u_i)\mathcal{S}(u_i, \theta_0)}{\sigma + \mu(u_i)}\right\| > t - \sigma/L_2\right), \quad (46)$$

for $t > \sigma/L_2$. Suppose we choose $\gamma = O(n^{-2/(8m+1)})$, $\sigma = O(\gamma)$, and $t = n^{-1/(16m+2)}$. Then combining (44) and (46) gives that for sufficiently large n we have

$$\mathbb{P}\left(\|\bar{x}_i - \mathcal{S}(u_i, \theta_0)\| > n^{-1/(16m+2)}\right) \leq c_6 \exp\left(-c_7 n^{1/2}\right), \quad (47)$$

where $c_6, c_7 > 0$ are constants. And so combining the union bound with (47) gives

$$\begin{aligned} \mathbb{P}\left(\max_{i \in [n]} \|\bar{x}_i - \mathcal{S}(u_i, \theta_0)\| > n^{-1/(16m+2)}\right) & \leq n \mathbb{P}\left(\|\bar{x}_i - \mathcal{S}(u_i, \theta_0)\| > n^{-1/(16m+2)}\right) \\ & \leq c_6 \exp\left(-c_7 n^{1/2} + \log n\right). \end{aligned} \quad (48)$$

The final implication of the result follows by noting that $n^{-2/(8m+1)} \rightarrow 0$ and $c_6 \exp(-c_7 n^{1/2} + \log n) \rightarrow 0$ as $n \rightarrow \infty$. \square

Before we present our algorithm, we need one more result that provides additional understanding for the semiparametric approach. Consider the following optimization problem

$$\text{ROBUST-IOP-SAA} \quad \min_{\theta} \max_{\epsilon \geq 0} \{Q_n(\theta; \epsilon) \mid \theta \in \Theta\},$$

PROPOSITION 10. *Suppose **A1**, **A2** and **R1** hold. Then the solution sets in θ of ROBUST-IOP-SAA and IOP-SAA are equivalent, and the optimal value of ROBUST-IOP-SAA occurs at $\epsilon = 0$.*

Proof. Let $\mathcal{C}_n(\theta, \epsilon)$ be the feasible set of R-DB-RISK-SAA. As shown in the proof for Proposition 8, the feasible set satisfies

$$\mathcal{C}_n(\theta, 0) \subseteq \mathcal{C}_n(\theta, \epsilon), \quad (49)$$

for all $\epsilon \geq 0$. As a result, we must have that $Q_n(\theta; 0) \geq Q_n(\theta; \epsilon)$ for all $\epsilon \geq 0$. This means that $\max_{\epsilon \geq 0} Q_n(\theta; \epsilon) = Q_n(\theta; 0)$. The result holds because $Q_n(\theta; 0) = Q_n(\theta)$ by definition. \square

Given the above relationship that the optimal value of ROBUST-IOP-SAA occurs at $\epsilon = 0$, we propose to solve the inverse optimization problem using the following formulation:

$$\begin{aligned} \text{SP-IOP-RISK-SAA} \quad & \hat{\theta}_n \in \arg \min \frac{1}{n} \sum_{i=1}^n \epsilon_i \\ & \text{s.t. } f(\bar{x}_i, u_i, \theta) - h(\lambda_i, u_i, \theta) \leq \epsilon_i, \quad \forall i \in [n] \\ & \lambda_i \geq 0, \quad \forall i \in [n] \end{aligned}$$

where the \bar{x}_i are as defined in (41). This is a convex optimization problem.

PROPOSITION 11. *Suppose **A1**–**A3** and **R1** hold. Then SP-IOP-RISK-SAA is a convex optimization problem.*

Proof. Since the Lagrangian dual function is defined as

$$h(\lambda, u, \theta) = \inf_x \left(f_0(x, u) + \sum_{j=1}^p \langle \theta \rangle_j f_j(x, u) + \sum_{j=1}^q \langle \lambda \rangle_j \langle g_0(x, u) \rangle_j \right), \quad (50)$$

it is the pointwise infimum of a family of affine functions of (λ, θ) and hence concave in (λ, θ) (Boyd and Vandenberghe 2009). Since $f(x, u, \theta)$ is affine in θ by **A3**, this means the constraint $f(\bar{x}_i, u_i, \theta) - h(\lambda_i, u_i, \theta) \leq \epsilon_i$ is convex in θ, ϵ_i . The objective function is linear, and so the entire optimization problem is convex. \square

We now have the elements to construct our semiparametric algorithm, which is a two-step approach. In the first step, we de-noise the y_i data using the L2NW estimator given in (41). And the second step is to solve SP-IOP-RISK-SAA. This approach is formally presented in Algorithm 2. Importantly, it can be shown that this semiparametric algorithm generates nearly-optimal solutions of IOP-SAA. This means the statistical consistency results in Section 3.3 apply to the solutions computed by this algorithm. In practice, the values of γ, σ can be chosen using standard approaches from statistics like cross-validation (Hastie et al. 2009).

Algorithm 2: Semiparametric Algorithm

Data: fixed $\gamma > 0$ and $\sigma > 0$

Result: estimate $\hat{\theta}_n$

- 1 **foreach** $i \in [n]$ **do**
 - 2 compute \bar{x}_i using (41);
 - 3 compute $\hat{\theta}_n$ using SP-IOP-RISK-SAA;
-

THEOREM 7. Suppose **A1–A3** and **R1–R4** and **IC** hold. If $\sigma = O(n^{-2/(8m+1)})$, $\lambda = O(\sigma)$, and $\hat{\theta}_n$ is computed using the semiparametric algorithm (i.e., Algorithm 2); then $\hat{\theta}_n$ is nearly-optimal for IOP-SAA in probability.

Proof. Note that $\min\{-h(\lambda, u, \theta) \mid \lambda \geq 0\} = -f(\mathcal{S}(u, \theta), u, \theta)$ by strong duality (which holds because of **A1, R1** (Bonnans and Shapiro 2000)). Next, consider the function

$$R(\theta) = \mathbb{E} \left(\min_{\lambda \geq 0} f(\mathcal{S}(u, \theta_0), u, \theta) - h(\lambda, u, \theta) \right) = \mathbb{E} \left(f(\mathcal{S}(u, \theta_0), u, \theta) - f(\mathcal{S}(u, \theta), u, \theta) \right), \quad (51)$$

its sample average approximation

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^n \left(\min_{\lambda_i \geq 0} f(\mathcal{S}(u_i, \theta_0), u_i, \theta) - h(\lambda_i, u_i, \theta) \right) = \frac{1}{n} \sum_{i=1}^n \left(f(\mathcal{S}(u_i, \theta_0), u_i, \theta) - f(\mathcal{S}(u_i, \theta), u_i, \theta) \right), \quad (52)$$

and its semiparametric approximation

$$\bar{R}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \left(\min_{\lambda_i \geq 0} f(\bar{x}_i, u_i, \theta) - h(\lambda_i, u_i, \theta) \right) = \frac{1}{n} \sum_{i=1}^n \left(f(\bar{x}_i, u_i, \theta) - f(\mathcal{S}(u_i, \theta), u_i, \theta) \right). \quad (53)$$

Note that $\min\{\bar{R}_n(\theta) \mid \theta \in \Theta\}$ is simply a reformulation of SP-IOP-RISK-SAA. Next, observe that $\mathbb{E}[f(\mathcal{S}(u, \theta_0), u, \theta) - f(\mathcal{S}(u, \theta), u, \theta) \mid u \in \mathcal{U}(\theta)] > 0$ since (i) $f(x, u, \theta)$ is twice continuously differentiable in x by **R3**, and (ii) $\text{dist}(\mathcal{S}(u, \theta), \mathcal{S}(u, \theta_0)) > 0$ for each $u \in \mathcal{U}(\theta)$ by **IC**. Consequently, we have $R(\theta) > 0$ for $\theta \in \Theta \setminus \theta_0$. As shown in the proof for Proposition 2, $\mathcal{S}(u, \theta)$ is continuous in θ . And so $R_n(\theta)$ and $\bar{R}_n(\theta)$ are continuous because (i) $f(x, u, \theta)$ is twice continuously differentiable in x, θ by **R3**.

Next, recall that \mathcal{U} is bounded by **R4**, Θ is bounded by **R2**, $f(x, u, \theta)$ is twice continuously differentiable in x, θ by **R3**, and the feasible set of FOP is absolutely bounded by **R1**. This means there exists $L > 0$ such that for all $\theta \in \Theta$ we have $\max_{i \in [n]} |f(\bar{x}_i, u_i, \theta) - f(\mathcal{S}(u_i, \theta_0), u_i, \theta)| \leq Ln^{-1/(18m)}$ whenever $\max_{i \in [n]} \|\bar{x}_i - \mathcal{S}(u_i, \theta_0)\| \leq n^{-1/(18m)}$ (which occurs with probability at least $1 - k_1 \exp(-k_2 n^{1/4})$ by Proposition 9). Thus, we have that $\sup_{\theta \in \Theta} |R_n(\theta) - \bar{R}_n(\theta)| \xrightarrow{p} 0$. Now consider any $\hat{\theta}_n \in \arg \min\{\bar{R}_n(\theta) \mid \theta \in \Theta\}$, and note that the estimate $\hat{\theta}_n$ returned by the semiparametric algorithm satisfies this property by construction. By definition we have $\bar{R}_n(\hat{\theta}_n) \leq \bar{R}_n(\theta_0)$, which can be rewritten as

$$R_n(\hat{\theta}_n) + \bar{R}_n(\hat{\theta}_n) - R_n(\hat{\theta}_n) \leq R_n(\theta_0) + \bar{R}_n(\theta_0) - R_n(\theta_0). \quad (54)$$

Thus, we have

$$R_n(\hat{\theta}_n) \leq R_n(\theta_0) + |\bar{R}_n(\hat{\theta}_n) - R_n(\hat{\theta}_n)| + |\bar{R}_n(\theta_0) - R_n(\theta_0)|. \quad (55)$$

We have thus shown all the conditions required to apply Theorem 5.14 of (van der Vaart 2000), which gives $\hat{\theta}_n \xrightarrow{P} \theta_0$. Now let $\bar{\theta}_n \in \arg \min\{Q_n(\theta) \mid \theta \in \Theta\}$. By Theorem 4, we have $\bar{\theta}_n \xrightarrow{P} \theta_0$. This means that $|\bar{\theta}_n - \hat{\theta}_n| \xrightarrow{P} 0$. \square

5. Numerical Experiments

We present numerical results that demonstrate the statistical consistency of our algorithms for inverse optimization with noisy data, and the results show our algorithms perform competitively against KKA (Keshavarz et al. 2011) and VIA (Bertsimas et al. 2014). We begin by conducting two types of tests using synthetic data. The first type is where the model is kept fixed and the number of data points increases, and the purpose is to display either estimation consistency or risk consistency of our algorithms. The second type is where the number of data points is kept fixed and the number of the parameters in the model increases, and the purpose is to display the feasibility of using our algorithms on large problems and to show the effect of data sampling and model complexity on statistical performance of our algorithms. Next, we apply our framework to a real data set in order to estimate a utility function that describes the tradeoff made between occupant comfort and the amount of energy consumption, when setting a thermostat temperature setpoint for air-conditioning.

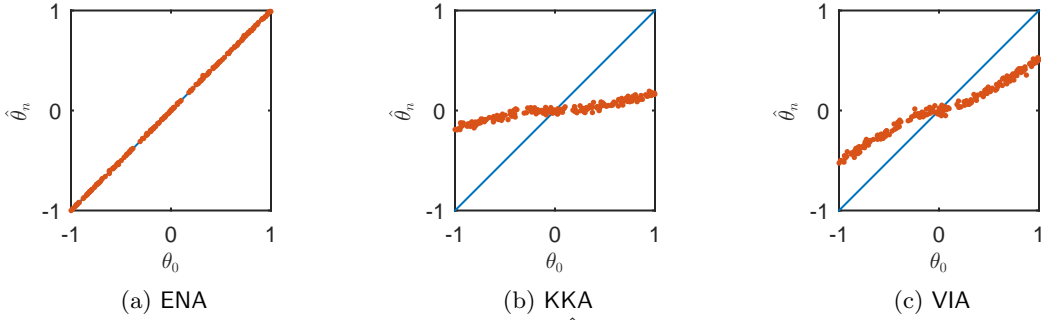
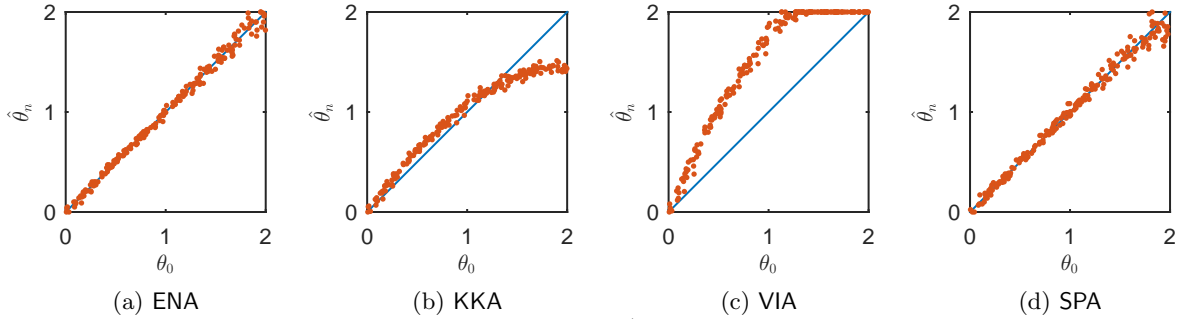
5.1. Synthetic Data and Enumeration Algorithm

In the first experiments, we generate data using a given FOP and then use the same set of equations in SAA-IOP. In other words, the first set of experiments are situations where the model whose parameters are being identified exactly matches the model that generates the data. As a result, this setting consists of situations where **IC** is satisfied. The first example is where: (i) FOP-A is $\min\{(\theta + u) \cdot x \mid x \in [-1, 1]\}$, (ii) u has a uniform distribution with support $[-1, 1]$, (iii) the measurement noise w has a normal distribution with zero mean and unit variance, (iv) the data is generated with $\theta_0 = 1$, and (v) the enumeration algorithm (i.e., Algorithm 1) was applied with $\epsilon = 0.001$, $\delta = 0.01$, and $\Theta = [-1, 1]$. The second example is where: (i) FOP-B is $\min\{x^2 - (\theta + u) \cdot x \mid x \in [0, 1]\}$, (ii) u has a uniform distribution with support $[0, 2]$, (iii) the measurement noise w has a normal distribution with zero mean and unit variance, (iv) the data is generated with $\theta_0 = \frac{1}{2}$, and (v) the enumeration algorithm (i.e., Algorithm 1) was applied with $\epsilon = 0$, $\delta = 0.01$, and $\Theta = [0, 2]$.

The results averaged over 100 repetitions of sampling $n \in \{10, 30, 50, 100, 300, 500, 1000\}$ data points and then estimating the parameter θ are summarized in Table 1. We label the enumeration algorithm (i.e., Algorithm 1) as **ENA** in the table. These results display estimation consistency

Table 1 Estimation Error $|\hat{\theta}_n - \theta_0|$ Using Different Algorithms

| | n | 10 | 30 | 50 | 100 | 300 | 500 | 1000 |
|--------------|-----|--------|--------|--------|--------|--------|--------|--------|
| Data: FOP-A | ENA | 0.2616 | 0.0926 | 0.0380 | 0.0211 | 0.0055 | 0.0030 | 0.0009 |
| Model: FOP-A | KKA | 0.8686 | 0.8293 | 0.8182 | 0.8257 | 0.8130 | 0.8231 | 0.8170 |
| | VIA | 0.5552 | 0.4976 | 0.4829 | 0.4887 | 0.4807 | 0.4846 | 0.4780 |
| Data: FOP-B | ENA | 0.4577 | 0.2481 | 0.1510 | 0.0501 | 0.0222 | 0.0123 | 0.0063 |
| Model: FOP-B | KKA | 0.5065 | 0.2281 | 0.1595 | 0.0751 | 0.0398 | 0.0342 | 0.0238 |
| | VIA | 0.9488 | 0.7051 | 0.6344 | 0.4284 | 0.3145 | 0.3810 | 0.2962 |

**Figure 1** Scatter plot comparing estimated parameter $\hat{\theta}_n$ versus true parameter θ_0 as computed by different algorithms at $n = 1,000$ when the data and model are both FOP-A.**Figure 2** Scatter plot comparing estimated parameter $\hat{\theta}_n$ versus true parameter θ_0 as computed by different algorithms at $n = 10,000$ when the data and model are both FOP-B.

of the enumeration algorithm since estimation error is decreasing to zero. To further illustrate estimation consistency, we conducted an experiment with the two examples above where the data was generated with a θ_0 that was randomly chosen from a uniform distribution with support $[-1, 1]$ and $[0, 2]$ for the first and second examples, respectively. A plot comparing the estimates $\hat{\theta}_n$ to the true parameter θ_0 for the first situation when $n = 1,000$ is shown in Figure 1, and a plot comparing the estimates $\hat{\theta}_n$ to the true parameter θ_0 for the second situation when $n = 10,000$ is shown in Figure 2. Consistent estimates should line up along the diagonal, and hence these plots demonstrate the estimation consistency (inconsistency) of the enumeration algorithm (KKA and VIA).

Table 2 Normalized Prediction Error $Q(\hat{\theta}_n) - \text{var}(w)$ Using Different Algorithms

| | | n | 10 | 30 | 50 | 100 | 300 | 500 | 1000 |
|-----------------------------|-----|--------|--------|--------|--------|--------|--------|--------|------|
| Data: FOP-C Model: FOP-B | ENA | 0.0216 | 0.0184 | 0.0162 | 0.0150 | 0.0065 | 0.0046 | 0.0017 | |
| | KKA | 0.0168 | 0.0124 | 0.0128 | 0.0151 | 0.0150 | 0.0150 | 0.0132 | |
| | VIA | 0.0249 | 0.0185 | 0.0196 | 0.0149 | 0.0089 | 0.0072 | 0.0042 | |
| Data: SQR-1 Model: FOP-B | ENA | 0.0294 | 0.0217 | 0.0152 | 0.0110 | 0.0073 | 0.0041 | 0.0024 | |
| | KKA | 0.0394 | 0.0389 | 0.0398 | 0.0440 | 0.0504 | 0.0525 | 0.0518 | |
| | VIA | 0.0343 | 0.0287 | 0.0243 | 0.0187 | 0.0122 | 0.0084 | 0.0072 | |

In the second set of experiments, we generate data using a given model that is different than the FOP used to formulate SAA-IOP. In other words, this set of experiments are situations where the model whose parameters are being identified does not match the model that generates the data. As a result, this setting consists of situations where **IC** is *not* satisfied. The first example is where: (i) the data is generated by FOP-C which is $\min\{\frac{3}{2} \cdot x^2 - (1+u) \cdot x \mid x \in [0, 1]\}$, (ii) the model estimated by IOP-SAA is FOP-B, (iii) u has a uniform distribution with support $[0, 5]$, (iv) the measurement noise w has a normal distribution with zero mean and unit variance, and (v) the enumeration algorithm (i.e., Algorithm 1) was applied with $\epsilon = 0$, $\delta = 0.01$, and $\Theta = [0, 2]$. The second example is where: (i) the data is generated by the statistical model SQR-1 given by $y_i = \min\{\max\{\sqrt{u_i}, 0\}, 1\} + w_i$, (ii) the model estimated by IOP-SAA is FOP-B, (iii) u has a uniform distribution with support $[0, 5]$, (iv) the measurement noise w has a normal distribution with zero mean and unit variance, and (v) the enumeration algorithm (i.e., Algorithm 1) was applied with $\epsilon = 0$, $\delta = 0.01$, and $\Theta = [0, 2]$.

The results averaged over 100 repetitions of sampling $n \in \{10, 30, 50, 100, 300, 500, 1000\}$ data points and then estimating the parameter θ are summarized in Table 2, and these results are normalized by subtracting $\text{var}(w)$. The reason for this normalization is that the prediction error $\mathbb{E}((y - \xi(u))^2)$ of the prediction $\xi(u)$ of the true model (either FOP-C or SQR-M, respectively) is $\text{var}(w)$ because $y = \xi(u) + w$ here. The enumeration algorithm has lower prediction error because it is risk consistent, whereas KKA and VIA are not risk consistent.

5.2. Synthetic Data and Semiparametric Algorithm

Next, we generate data using a given FOP and then use the same equations in SAA-IOP. These experiments are situations where the model whose parameters are being identified exactly matches the model that generates the data. As a result, this setting consists of situations where **IC** is satisfied. The first example is where: (i) FOP-D is $\min\{x'x - (\theta + u)'x \mid x \in [0, 1]^p\}$, (ii) u has a uniform distribution with support $[0, 2]^p$, (iii) the measurement noise w has a jointly Gaussian distribution with zero mean and identity covariance, (iv) the data is generated with $p = 10$ and $\theta_0 \in \mathbb{R}^p$ such that $\langle \theta_0 \rangle_k = \frac{1}{2}$ for all $k \in [p]$, and (v) the semiparametric algorithm (i.e., Algorithm 2)

was applied with γ, σ chosen using cross-validation (Hastie et al. 2009) and $\Theta = [0, 2]$. The second example is where: (i) FOP-E is

$$\min\{-\sum_{k=1}^p \langle \theta \rangle_k \cdot \log(\langle x \rangle_k + \langle u \rangle_k) - \log(\langle x \rangle_{p+1} + \langle u \rangle_{p+1}) \mid \langle x \rangle_k \geq 0, \sum_{k=1}^{p+1} \langle x \rangle_k = 1\}, \quad (56)$$

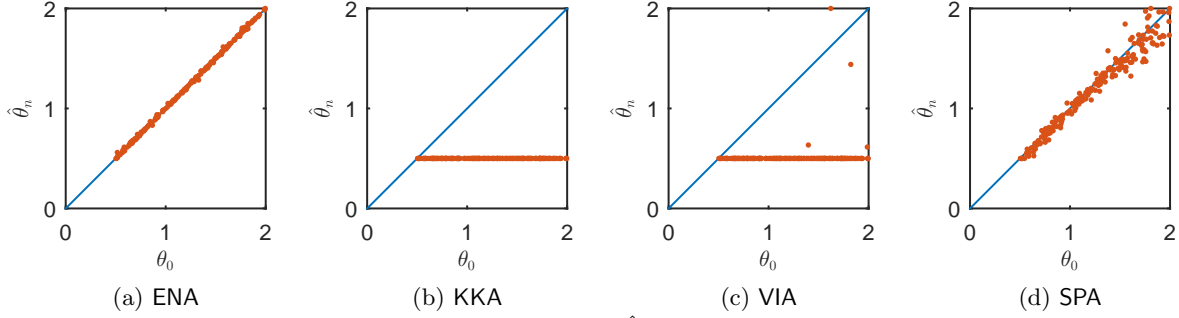
(ii) u has a uniform distribution with support $[1, 2]^{p+1}$, (iii) the measurement noise w has a jointly Gaussian distribution with zero mean and identity covariance, (iv) the data is generated with $p = 10$ and $\theta_0 \in \mathbb{R}^p$ such that $\langle \theta_0 \rangle_k = 1$ for all $k \in [p]$, and (v) a modified version of the semiparametric algorithm (i.e., Algorithm 2) was applied with γ, σ chosen using cross-validation (Hastie et al. 2009) and $\Theta = [\frac{1}{2}, 2]$. The modification to Algorithm 2 is we calculate $\tilde{x}_i = \min_x \{\|\bar{x}_i - x\| \mid \langle x \rangle_k \geq 0\}$ and then compute $\hat{\theta}_n$ using SP-IOP-RISK-SAA with the \tilde{x}_i replacing the \bar{x}_i . The \tilde{x}_i are the projection of the \bar{x}_i onto the nonnegative orthant, and it turns out this projection does not affect our theoretical results. In particular, a short proof using the continuous mapping theorem (van der Vaart 2000) and the boundedness of the feasible set in **R1** gives that $\max_{i \in [n]} \|\tilde{x}_i - \mathcal{S}(u_i, \theta_0)\| \xrightarrow{p} 0$. The projection is needed for this particular example because otherwise we would have logarithms of negative numbers, which are complex-valued. More generally, a projection of \bar{x}_i onto the feasible set of FOP will not affect our theoretical results, and can be added as a step in our semiparametric algorithm.

The results averaged over 100 repetitions of sampling $n \in \{10, 30, 50, 100, 300, 500, 1000\}$ data points and then estimating the parameter θ are summarized in Table 3. We label the semiparametric algorithm (i.e., Algorithm 2) as SPA in the table. These results display estimation consistency of the semiparametric algorithm since it has lower estimation error as the data increases. To further illustrate estimation consistency, we conducted an experiment with the two situations above where the data was generated with $p = 1$ and a θ_0 that was randomly chosen from a uniform distribution with support $[0, 1]$ and $[\frac{1}{2}, 2]$ for the first and second situations, respectively. A plot comparing the estimates $\hat{\theta}_n$ to the true parameter θ_0 for the first situation when $n = 1,000$ is shown in Figure 2, and a plot comparing the estimates $\hat{\theta}_n$ to the true parameter θ_0 for the second situation when $n = 1,000$ is shown in Figure 3. Consistent estimates should line up along the diagonal, and hence these plots demonstrate the estimation consistency (inconsistency) of the semiparametric algorithm (KKA and VIA). It is worth comparing the results of the semiparametric and enumeration algorithms. As mentioned above, the semiparametric algorithm will generally have higher estimation error – this can be observed in these plots because the semiparametric algorithm estimates have a larger variation about the diagonal than the estimates of the enumeration algorithm.

In the second set of experiments, we generate data using a given model that is different than the FOP used to formulate SAA-IOP. In other words, this set of experiments are situations where the model whose parameters are being identified does not match the model that generates the

Table 3 Estimation Error $\|\hat{\theta}_n - \theta_0\|$ Using Different Algorithms

| | n | 10 | 30 | 50 | 100 | 300 | 500 | 1000 |
|-----------------------------|-----|--------|--------|--------|--------|--------|--------|--------|
| Data: FOP-D Model: FOP-D | SPA | 2.4618 | 1.7025 | 1.2543 | 0.8535 | 0.4754 | 0.3750 | 0.2573 |
| | KKA | 2.2569 | 1.5513 | 1.2229 | 0.9281 | 0.6107 | 0.5435 | 0.4447 |
| | VIA | 3.3829 | 3.2603 | 3.1937 | 3.1501 | 3.0292 | 3.0324 | 2.9208 |
| Data: FOP-E Model: FOP-E | SPA | 0.9189 | 0.7982 | 0.7500 | 0.7487 | 0.6639 | 0.6070 | 0.5783 |
| | KKA | 1.6687 | 1.5850 | 1.5813 | 1.5865 | 1.5828 | 1.5806 | 1.5811 |
| | VIA | 1.9299 | 1.6781 | 1.6826 | 1.6132 | 1.6001 | 1.5973 | 1.5843 |

**Figure 3** Scatter plot comparing estimated parameter $\hat{\theta}_n$ versus true parameter θ_0 as computed by different algorithms at $n = 1,000$ when the data and model are both FOP-E with $p = 1$.

data. As a result, this setting consists of situations where **IC** is *not* satisfied. The first example is where: (i) the data is generated by FOP-F which is $\min\{\frac{3}{2} \cdot x'x' - (1+u)'x \mid x \in [0,1]^{10}\}$, (ii) the model estimated by IOP-SAA is FOP-D with $p = 10$, (iii) u has a uniform distribution with support $[0,5]^{10}$, (iv) the measurement noise w has a jointly Gaussian distribution with zero mean and identity covariance, and (v) the semiparametric algorithm (i.e., Algorithm 2) was applied with γ, σ chosen using cross-validation (Hastie et al. 2009) and $\Theta = [0, 2]$. The second example is where: (i) the data is generated by the statistical model SQR-P given by $y_i = \min\{\max\{\sqrt{u_i}, 0\}, 1\} + w_i$, (ii) the model estimated by IOP-SAA is FOP-D with $p = 10$, (iii) u has a uniform distribution with support $[0,5]^{10}$, (iv) the measurement noise w has a jointly Gaussian distribution with zero mean and identity covariance, and (v) the semiparametric algorithm (i.e., Algorithm 2) was applied with γ, σ chosen using cross-validation (Hastie et al. 2009) and $\Theta = [0, 2]$.

The results averaged over 100 repetitions of sampling $n \in \{10, 30, 50, 100, 300, 500, 1000\}$ data points and then estimating the parameter θ are summarized in Table 4, and these results are normalized by subtracting $\mathbb{E}(w'w)$. The reason for this normalization is that the prediction error $\mathbb{E}(\|y - \xi(u)\|^2)$ of the prediction $\xi(u)$ of the true model (either FOP-C or SQR-M, respectively) is $\mathbb{E}(w'w)$ because $y = \xi(u) + w$ here. The enumeration algorithm has lower prediction error because it is risk consistent, whereas KKA and VIA are not risk consistent.

In the third set of experiments, we generate data using the previous four settings. The difference in this set of experiments is that we fix $n = 1000$ and vary $p \in \{1, 3, 5, 10, 30\}$. The results when

Table 4 Normalized Prediction Error $Q(\hat{\theta}_n) - \mathbb{E}(w'w)$ Using Different Algorithms

| | | n | 10 | 30 | 50 | 100 | 300 | 500 | 1000 |
|-----------------------------|-----|-----|--------|--------|--------|--------|--------|--------|--------|
| Data: FOP-C Model: FOP-D | SPA | | 0.2319 | 0.1972 | 0.1744 | 0.1501 | 0.1029 | 0.0844 | 0.0529 |
| | KKA | | 0.1584 | 0.1308 | 0.1314 | 0.1349 | 0.1452 | 0.1497 | 0.1481 |
| | VIA | | 0.3438 | 0.3407 | 0.3360 | 0.3205 | 0.2950 | 0.2816 | 0.2811 |
| Data: SQR-M Model: FOP-D | SPA | | 0.4180 | 0.3497 | 0.3195 | 0.2470 | 0.1572 | 0.0998 | 0.0658 |
| | KKA | | 0.3645 | 0.3885 | 0.3987 | 0.4537 | 0.5115 | 0.5114 | 0.5214 |
| | VIA | | 0.3468 | 0.2784 | 0.2737 | 0.2524 | 0.2405 | 0.2458 | 0.2599 |

Table 5 Estimation Error $\|\hat{\theta}_n - \theta_0\|$ Using Different Algorithms

| | | p | 1 | 3 | 5 | 10 | 30 |
|-----------------------------|-----|-----|--------|--------|--------|--------|--------|
| Data: FOP-D Model: FOP-D | SPA | | 0.0601 | 0.1464 | 0.1907 | 0.2794 | 0.4701 |
| | KKA | | 0.1178 | 0.2349 | 0.3038 | 0.4619 | 0.7978 |
| | VIA | | 0.4943 | 1.2254 | 1.8099 | 2.9522 | 5.7737 |
| Data: FOP-E Model: FOP-E | SPA | | 0.0251 | 0.1258 | 0.2571 | 0.5890 | 0.5576 |
| | KKA | | 0.5000 | 0.8660 | 1.1174 | 1.5804 | 2.7377 |
| | VIA | | 0.5000 | 0.8691 | 1.1231 | 1.5966 | 2.7628 |

Table 6 Normalized Prediction Error $Q(\hat{\theta}_n) - \mathbb{E}(w'w)$ Using Different Algorithms

| | | p | 1 | 3 | 5 | 10 | 30 |
|-----------------------------|-----|-----|--------|--------|--------|--------|--------|
| Data: FOP-C Model: FOP-D | SPA | | 0.0064 | 0.0171 | 0.0403 | 0.0628 | 0.2048 |
| | KKA | | 0.0538 | 0.1553 | 0.2619 | 0.5252 | 1.5712 |
| | VIA | | 0.0078 | 0.0175 | 0.0745 | 0.2602 | 0.9654 |
| Data: SQR-M Model: FOP-D | SPA | | 0.0056 | 0.0194 | 0.0319 | 0.0606 | 0.1568 |
| | KKA | | 0.0148 | 0.0471 | 0.0761 | 0.1523 | 0.4394 |
| | VIA | | 0.0055 | 0.0273 | 0.0821 | 0.2848 | 1.2896 |

the data/model are given by FOP-D/FOP-D and FOP-E/FOP-E, averaged over 100 repetitions and then estimating the parameter θ , are summarized in Table 5. These results show that the semiparametric algorithm has lower estimation error than KKA and VIA on these examples. The results when the data/model are given by FOP-C/FOP-B and SQR-M/FOP-B, averaged over 100 repetitions and then estimating the parameter θ , are summarized in Table 6. These results show that the semiparametric algorithm has lower prediction error than KKA and VIA on these examples.

5.3. Empirical Data: Estimating an Energy-Comfort Utility Function

We next apply our inverse optimization framework to the problem of estimating a utility function that describes the tradeoff made between occupant comfort and the amount of energy consumption, when setting a thermostat temperature setpoint for air-conditioning. The data we use is collected from Sutardja Dai Hall on the Berkeley campus, which was used as part of the BRITE-S testbed in our past experiments (Aswani et al. 2012a,b,c) concerning robust learning-based optimization (Aswani et al. 2013) of heating, ventilation, and air-conditioning (HVAC) systems. Specifically, this building is equipped with a commercial web application (Building Robotics 2016) that allows

Table 7 Prediction Error $Q(\hat{\theta}_n)$ Using Different Algorithms

| | | n | 10 | 30 | 50 | 100 | 300 | 500 | 1000 |
|-----------------------------|-----|--------|--------|--------|--------|--------|--------|--------|------|
| Data: SDH-E Model: FOP-S | ENA | 1.3656 | 1.3308 | 1.3255 | 1.3169 | 1.3112 | 1.3099 | 1.3090 | |
| | KKA | 2.2439 | 2.2528 | 2.2508 | 2.2351 | 2.2225 | 2.2220 | 2.2200 | |
| | VIA | 2.2975 | 2.2538 | 2.2472 | 2.2277 | 2.2163 | 2.2166 | 2.2138 | |

occupants to change the thermostat temperature setpoints in real-time, and so the setpoints are changed throughout the year by occupants in response to factors like the outside weather.

When a room is being cooled, a lower temperature setpoint requires increased energy consumption since the air-conditioner must provide more cold air; however, the purpose of air-conditioning is to improve comfort by lowering the room temperature. And so individuals must tradeoff comfort and energy consumption when choosing the setpoint. A simplified utility function model (expressed as minimization of the negative of the utility function) that captures this tradeoff is FOP-S:

$$\min_x \{ \langle \theta \rangle_1 \cdot (x - 76)^2 + (x - \langle \theta \rangle_2 - u)^2 \mid x \in [70, 76] \}, \quad (57)$$

where $x \in \mathbb{R}$ is the thermostat temperature setpoint in units of degrees Fahrenheit ($^{\circ}\text{F}$), and $u \in \mathbb{R}$ is the current outside temperature in degrees Fahrenheit ($^{\circ}\text{F}$). The term $(x - \langle \theta \rangle_2 - u)^2$ indicates a preference for a temperature setpoint that is a fixed amount $\langle \theta \rangle_2$ above the outside temperature u (i.e., the preferred temperature is $\langle \theta \rangle_2 + u$), and the reason for this term is that individuals prefer a higher indoor temperature as the outside temperature increases (ASHRAE 2013). The term $\langle \theta \rangle_1 \cdot (x - 76)^2$ indicates a preference for a higher setpoint because of energy considerations, and the number 76 is used because 76°F – 78°F is a relatively high setpoint temperature that is often recommended for saving energy. The parameter $\langle \theta \rangle_1$ quantifies the tradeoff between the preference for a higher setpoint to save energy versus the desired indoor temperature $\langle \theta \rangle_2 + u$. Lastly, the constraints $x \in [70, 76]$ indicate observed setpoint limits.

The results averaged over 100 repetitions of sampling $n \in \{10, 30, 50, 100, 300, 500, 1000\}$ data points and then estimating the parameters θ are summarized in Table 7. The data set (which we label SDH-E in the table) used consists of outside temperature measurements (i.e., the u variable) and the chosen temperature set point (i.e., the x variable) of a single thermostat in Sutardja Dai Hall. In each repetition, the full data set was randomly split into a 1,000 point *training* data set and a 14,500 point *testing* data set. The n data points were randomly chosen from the training data set, and the prediction error of the estimated parameters were computed using the testing data set. To evaluate the statistical significance of the computed results, a bootstrap hypothesis test (Efron and Tibshirani 1994) was conducted. The computed p -value was less than 0.01, which indicates that the improved performance of the enumeration algorithm is statistically significant.

6. Conclusion

We developed and analyzed a formulation for inverse optimization in the setting where noisy measurements of the optimal points of a convex optimization problem are available. Our approach requires solving a bilevel program, and we defined a new duality-based reformulation to convert this bilevel program into a single level program. We showed that existing heuristics for solving the problem of inverse optimization with noisy data are statistically inconsistent, whereas our formulation as a bilevel program leads to statistical consistency. Though our formulation is NP-hard to solve, we provided two numerical algorithms for solving our formulation and then demonstrated the improved estimates generated by our approaches through a series of numerical experiments involving both synthetic and empirical data.

Acknowledgments

The authors gratefully acknowledge the support of NSF Award CMMI-1450963 and an NSERC Postgraduate Scholarship.

Bibliography

- Ahuja, Ravindra K, James B Orlin. 2001. Inverse optimization. *Operations Research* **49**(5) 771–783.
- ASHRAE. 2013. *ANSI/ASHRAE Standard 55-2013: Thermal Environmental Conditions for Human Occupancy*. ASHRAE.
- Aswani, A. 2015. Low-rank approximation and completion of positive tensors URL <http://arxiv.org/abs/1412.0620>. ArXiv:1412.0620.
- Aswani, A., H. Gonzalez, S. Sastry, C. Tomlin. 2013. Provably safe and robust learning-based model predictive control. *Automatica* **49**(5) 1216–1226.
- Aswani, A., P. Kaminsky, Y. Mintz, E. Flowers, Y. Fukuoka. 2016. Predictive modeling of behavior in weight loss interventions Submitted.
- Aswani, A., N. Master, J. Taneja, A. Krioukov, D. Culler, C. Tomlin. 2012a. Energy-efficient building HVAC control using hybrid system LBMPC. *IFAC Conference on Nonlinear Model Predictive Control*.
- Aswani, A., N. Master, J. Taneja, A. Krioukov, D. Culler, C. Tomlin. 2012b. Quantitative methods for comparing different HVAC control schemes. *International Conference on Performance Evaluation Methodologies and Tools*.
- Aswani, A., N. Master, J. Taneja, V. Smith, A. Krioukov, D. Culler, C. Tomlin. 2012c. Identifying models of HVAC systems using semi-parametric regression. *American Control Conference*.
- Aswani, A., C. Tomlin. 2012. Incentive design for efficient building quality of service. *Allerton Conference on Communication, Control, and Computing*. 90–97.

-
- Audet, C., P. Hansen, B. Jaumard, G. Savard. 1997. Links between linear bilevel and mixed 0–1 programming problems. *Journal of Optimization Theory and Applications* **93**(2) 273–300. doi:10.1023/A:1022645805569. URL <http://dx.doi.org/10.1023/A:1022645805569>.
- Bajari, Patrick, C Lanier Benkard, Jonathan Levin. 2007. Estimating dynamic models of imperfect competition. *Econometrica* **75**(5) 1331–1370.
- Bartlett, P., S. Mendelson. 2002. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* .
- Beil, Damian R, Lawrence M Wein. 2003. An inverse-optimization-based auction mechanism to support a multiattribute rfq process. *Management Science* **49**(11) 1529–1545.
- Berge, Claude. 1963. *Topological Spaces: including a treatment of multi-valued functions, vector spaces, and convexity*. Courier Dover Publications.
- Bertsimas, Dimitris, Vishal Gupta, Ioannis Ch Paschalidis. 2012. Inverse optimization: a new perspective on the black-litterman model. *Operations research* **60**(6) 1389–1403.
- Bertsimas, Dimitris, Vishal Gupta, Ioannis Ch Paschalidis. 2014. Data-driven estimation in equilibrium using inverse optimization. *Mathematical Programming Series A* .
- Bickel, P., K. Doksum. 2006. *Mathematical Statistics: Basic Ideas And Selected Topics*, vol. 1. 2nd ed. Pearson Prentice Hall.
- Bonnans, J., A. Shapiro. 2000. *Perturbation Analysis of Optimization Problems*. Springer.
- Bonnans, J Frederic, Alexander D Ioffe. 1995. Quadratic growth and stability in convex programming problems with multiple solutions. *J. Convex Anal* **2**(1-2) 41–57.
- Bonnans, JF. 1992. Directional derivatives of optimal solutions in smooth nonlinear programming. *Journal of optimization theory and applications* **73**(1) 27–45.
- Boyd, Stephen, Lieven Vandenbergh. 2009. *Convex optimization*. Cambridge university press.
- Building Robotics. 2016. Comfy. URL <https://gocomfy.com>.
- Burton, Didier, Ph L Toint. 1992. On an instance of the inverse shortest paths problem. *Mathematical Programming* **53**(1-3) 45–61.
- Carr, Scott, William Lovejoy. 2000. The inverse newsvendor problem: Choosing an optimal demand portfolio for capacitated resources. *Management Science* **46**(7) 912–927.
- Chan, Timothy CY, Tim Craig, Taewoo Lee, Michael B Sharpe. 2014. Generalized inverse multiobjective optimization with application to cancer therapy. *Operations Research* .
- Chatterjee, Sourav. 2014. A new perspective on least squares under convex constraint. *Ann. Statist.* **42**(6) 2340–2381. doi:10.1214/14-AOS1254. URL <http://dx.doi.org/10.1214/14-AOS1254>.
- Dempe, Stephan, Vyacheslav Kalashnikov, Gerardo Pérez-Valdés, Nataliya Kalashnikova. 2015. *Bilevel Programming Problems*. Springer.

-
- Efron, B., R.J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis.
- Erkin, Zeynep, Matthew D Bailey, Lisa M Maillart, Andrew J Schaefer, Mark S Roberts. 2010. Eliciting patients' revealed preferences: An inverse markov decision process approach. *Decision Analysis* **7**(4) 358–365.
- Faragó, András, Áron Szentesi, Balázs Szviatovszki. 2003. Inverse optimization in high-speed networks. *Discrete Applied Mathematics* **129**(1) 83–98.
- Greenshtein, E., Y. Ritov. 2004. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10**(6) 971–988. doi:10.3150/bj/1106314846. URL <http://dx.doi.org/10.3150/bj/1106314846>.
- Hastie, T., R. Tibshirani, J. Friedman. 2009. *The Elements of Statistical Learning*. 2nd ed. Springer-Verlag.
- Haviv, Ishay, Oded Regev. 2012. Tensor-based hardness of the shortest vector problem to within almost polynomial factors. *Theory of Computing* **8**(1) 513–531.
- Heuberger, Clemens. 2004. Inverse combinatorial optimization: A survey on problems, methods, and results. *Journal of Combinatorial Optimization* **8**(3) 329–361.
- Hillar, C., L.-H. Lim. 2013. Most tensor problems are np-hard. *J. ACM* **60**(6) 45:1–45:39. doi:10.1145/2512329. URL <http://doi.acm.org/10.1145/2512329>.
- Hochbaum, Dorit S. 2003. Efficient algorithms for the inverse spanning-tree problem. *Operations Research* **51**(5) 785–797.
- Iyengar, Garud, Wanmo Kang. 2005. Inverse conic programming with applications. *Operations Research Letters* **33**(3) 319–330.
- Jennrich, Robert I. 1969. Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics* 633–643.
- José Fortuny-Amat, Bruce McCarl. 1981. A representation and economic interpretation of a two-level programming problem. *The Journal of the Operational Research Society* **32**(9) 783–792.
- Keshavarz, Arezou, Yang Wang, Stephen Boyd. 2011. Imputing a convex objective function. *Intelligent Control (ISIC), 2011 IEEE International Symposium on*. IEEE, 613–619.
- Ratliff, Lillian J, Roy Dong, Henrik Ohlsson, S Shankar Sastry. 2014. Incentive design and utility learning via energy disaggregation. *Proceedings of the 19th IFAC World Congress*. 3158–3163.
- Rockafellar, R Tyrrell, Roger J-B Wets. 1998. *Variational analysis*, vol. 317. Springer.
- Saez-Gallego, J., J. M. Morales, M. Zugno, H. Madsen. 2016. A data-driven bidding model for a cluster of price-responsive consumers of electricity. *IEEE Transactions on Power Systems* **PP**(99) 1–11. doi:10.1109/TPWRS.2016.2530843.
- Schaefer, Andrew J. 2009. Inverse integer programming. *Optimization Letters* **3**(4) 483–489.

- Tao, T. 2012. *Topics in Random Matrix Theory*. Graduate studies in mathematics, American Mathematical Society.
- Troutt, Marvin D, Wan-Kai Pang, Shui-Hung Hou. 2006. Behavioral estimation of mathematical programming objective function coefficients. *Management science* **52**(3) 422–434.
- Tversky, Amos, Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* **211**(4481) 453–458.
- van der Vaart, A.W. 2000. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.
- Vershynin, Roman. 2012. *Compressed Sensing*, chap. Introduction to the non-asymptotic analysis of random matrices. Cambridge University Press, 210–268.
- Wald, Abraham. 1949. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20**(4) 595–601. doi:10.1214/aoms/1177729952. URL <http://dx.doi.org/10.1214/aoms/1177729952>.
- Wang, Lizhi. 2009. Cutting plane algorithms for the inverse mixed integer linear programming problem. *Operations Research Letters* **37**(2) 114–116.
- Zhang, Jianzhong, Zhenhong Liu. 1996. Calculating some inverse linear programming problems. *Journal of Computational and Applied Mathematics* **72**(2) 261–273.
- Zhang, Jianzhong, Chengxian Xu. 2010. Inverse optimization for linearly constrained convex separable programming problems. *European Journal of Operational Research* **200**(3) 671–679.

Appendix

A. Lemmas

LEMMA 1. Suppose **R4** holds. Then for $t > c_1 \cdot \gamma$ we have

$$\mathbb{P}\left(\left|\gamma^{-m} \cdot \frac{1}{n} \sum_{j=1}^n K\left(\frac{u_j - u_i}{\gamma}\right) - \mu(u_i)\right| > t\right) \leq 2 \exp\left(-2c_2 n \gamma^{2m} \cdot (t - c_1 \cdot \gamma)^2\right), \quad (58)$$

where $c_1, c_2 > 0$ are constants.

Proof. Recall $\mu(u)$ is the probability density function of u , and note that

$$\begin{aligned} \left|\mu(u_i) - \mathbb{E}\left[\gamma^{-m} K\left(\frac{u - u_i}{\gamma}\right) \middle| u_i\right]\right| &= \left|\mu(u_i) - \gamma^{-m} \int_{\mathbb{R}^m} K\left(\frac{u - u_i}{\gamma}\right) \mu(u) du\right| \\ &= \left|\mu(u_i) - \gamma^{-m} \int_{\mathbb{R}^m} K(s) \mu(u_i + \gamma s) \gamma^m ds\right| \\ &= \left|\mu(u_i) - \int_{\mathbb{R}^m} K(s) (\mu(u_i) + \gamma \nabla \mu(u_i + \beta \gamma s)^T s) ds\right| \\ &= \left|\int_{\mathbb{R}^m} K(s) \nabla \mu(u_i + \beta \gamma s)^T s ds\right| \cdot \gamma \\ &\leq c_1 \cdot \gamma, \end{aligned} \quad (59)$$

where the second line follows from a change of variables $s = (u - u_i)/\gamma$, the third line follows from the multivariate form of Taylor's Theorem with some $\beta \in [0, 1]$, the fourth line follows because a Kernel function has the property $\int K(u) du = 1$, and the fifth line follows by setting $c_1 = \max_{u \in \mathcal{U}} \left|\int_{\mathbb{R}^m} K(s) \nabla \mu(u)^T s ds\right|$. Note this c_1 term is finite because (i) a kernel function has the property that its support is finite (i.e., $K(u) = 0$ for $\|u\| > 1$), and (ii) $\mu(u)$ is a continuously differentiable probability density function by **R4**. Next, note that by Hoeffding's inequality (Vershynin 2012) we have for $t > 0$ that

$$\mathbb{P}\left(\left|\gamma^{-m} \cdot \frac{1}{n} \sum_{j=1}^n K\left(\frac{u_j - u_i}{\gamma}\right) - \mathbb{E}\left[\gamma^{-m} K\left(\frac{u - u_i}{\gamma}\right) \middle| u_i\right]\right| > t\right) \leq 2 \exp\left(-2c_2 n \gamma^{2m} t^2\right), \quad (60)$$

where $c_2 = (\max_u K(u))^2$. Combining (59) and (60) gives the desired result. \square

LEMMA 2. Suppose **A1** and **R1–R4** hold. Then for $t > c_3 \cdot \gamma^{1/2} + c_4 \cdot \gamma$ we have

$$\mathbb{P}\left(\left\|\gamma^{-m} \cdot \frac{1}{n} \sum_{j=1}^n y_j \cdot K\left(\frac{u_j - u_i}{\gamma}\right) - \mu(u_i) \mathcal{S}(u_i, \theta_0)\right\| > t\right) \leq 2 \exp\left(-2c_5 n \gamma^{2m} \cdot (t - c_3 \cdot \gamma^{1/2} - c_4 \cdot \gamma)\right). \quad (61)$$

where $c_3, c_4, c_5 > 0$ are constants.

Proof. First, note that $\mathcal{S}(u, \theta)$ consists of a single point from the strict convexity assumption in **R3**. Next, note that having **A1** and **R1–R4** means that Proposition 4.41 of (Bonnans and Shapiro 2000) holds: This means for $\gamma > 0$ sufficiently small we have

$$\|\mathcal{S}(u, \theta_0) - \mathcal{S}(u_i, \theta_0)\| \leq \alpha \cdot \gamma^{1/2}, \quad (62)$$

where $\alpha > 0$ is a constant, whenever $\|u - u_i\| \leq \gamma$. Next, recall that y_i conditioned on u_i has distribution $\mathcal{S}(u_i, \theta_0) + w_i$ under **IC**. Moreover, we have

$$\mathbb{E}\left[\gamma^{-m} y K\left(\frac{u - u_i}{\gamma}\right) \middle| u_i\right] = \mathbb{E}\left[\gamma^{-m} \mathcal{S}(u, \theta_0) K\left(\frac{u - u_i}{\gamma}\right) \middle| u_i\right], \quad (63)$$

since $\mathbb{E}(w_i) = 0$ and w_i is independent of u_i . Thus, we have

$$\begin{aligned}
& \left\| \mu(u_i) \mathcal{S}(u_i, \theta_0) - \mathbb{E} \left[\gamma^{-m} y K \left(\frac{u - u_i}{\gamma} \right) \middle| u_i \right] \right\| \\
&= \left\| \mu(u_i) \mathcal{S}(u_i, \theta_0) - \gamma^{-m} \int_{\mathbb{R}^m} K \left(\frac{u - u_i}{\gamma} \right) \mu(u) \mathcal{S}(u, \theta_0) du \right\| \\
&= \left\| \mu(u_i) \mathcal{S}(u_i, \theta_0) - \gamma^{-m} \int_{\mathbb{R}^m} K(s) \mu(u_i + \gamma s) \mathcal{S}(u_i + \gamma s, \theta_0) \gamma^m ds \right\| \\
&= \left\| \mu(u_i) \mathcal{S}(u_i, \theta_0) - \int_{\mathbb{R}^m} K(s) (\mu(u_i) + \gamma \nabla \mu(u_i + \beta \gamma s)^T s) (\mathcal{S}(u_i, \theta_0) + \right. \\
&\quad \left. \mathcal{S}(u_i + \gamma s, \theta_0) - \mathcal{S}(u_i, \theta_0)) ds \right\| \\
&= \left\| \int_{\mathbb{R}^m} K(s) \mu(u_i) (\mathcal{S}(u_i + \gamma s, \theta_0) - \mathcal{S}(u_i, \theta_0)) ds + \int_{\mathbb{R}^m} K(s) \gamma \nabla \mu(u_i + \beta \gamma s)^T s \mathcal{S}(u, \theta_0) ds \right\| \\
&\leq c_3 \cdot \gamma^{1/2} + c_4 \cdot \gamma,
\end{aligned} \tag{64}$$

where the second line follows from a change of variables $s = (u - u_i)/\gamma$, the third line follows from the multi-variate form of Taylor's Theorem with some $\beta \in [0, 1]$, the fourth line follows because a Kernel function has the property $\int K(u) du = 1$, and the fifth line follows from (62) and by setting $c_3 = \alpha \cdot \max_{u \in \mathcal{U}} |\int_{\mathbb{R}^m} K(s) \mu(u) ds|$ and $c_4 = \max_{u \in \mathcal{U}} (|\int_{\mathbb{R}^m} K(s) \nabla \mu(u)^T s ds| \cdot \|\mathcal{S}(u, \theta_0)\|)$. Note the c_3, c_4 terms are finite because (i) a kernel function has the property that its support is finite (i.e., $K(u) = 0$ for $\|u\| > 1$), (ii) $\mu(u)$ is a continuously differentiable probability density function by **R4**, and (iii) $\mathcal{S}(u, \theta_0)$ is bounded by **R1**. Next, note that y is a sub-exponential random variable (Vershynin 2012) since (i) $\mathcal{S}(u, \theta_0)$ is a bounded random variable by **R1**, and (ii) w is sub-exponential by **R4**. Hence, by Hoeffding's inequality for sub-exponential random variables (Vershynin 2012) we have for $t > 0$ that

$$\mathbb{P} \left(\left\| \gamma^{-m} \cdot \frac{1}{n} \sum_{j=1}^n y_j \cdot K \left(\frac{u_j - u_i}{\gamma} \right) - \mathbb{E} \left[\gamma^{-m} y K \left(\frac{u - u_i}{\gamma} \right) \middle| u_i \right] \right\| > t \right) \leq 2 \exp \left(-2c_5 n \gamma^{2m} t \right), \tag{65}$$

for some $c_5 > 0$. Combining (64) and (65) gives the desired result. \square

B. Identifiability in Inverse Optimization

Estimation consistency in any statistical setting (including inverse optimization with noisy data) requires that an identifiability condition holds, and such identifiability conditions can be stated under a variety of different mathematical formulations (Wald 1949, Jennrich 1969, Bartlett and Mendelson 2002, Greenshtein and Ritov 2004, Bickel and Doksum 2006, Chatterjee 2014, Aswani 2015). The intuition for these different formulations is the same: Essentially, an identifiability condition states that the output of the model is different for two distinct sets of model parameters. It is important to note that identifiability is a statistical property of the model and the error metric used. Consequently, it is possible for an estimator to be statistically inconsistent, even when an identifiability condition holds (see for instance Proposition proposition:estincon). In the context of inverse optimization with noisy data, we define an identifiability condition **IC**.

Showing that **IC** holds is complicated by the presence of constraints in FOP. To illustrate this, consider two related instances of FOP with $x \in \mathbb{R}$ and $\theta \in [0, 2]$. The first $\min(x - \theta)^2$ is FOP-I, and the second $\min\{(x - \theta)^2 \mid x \leq 1\}$ is FOP-II. Since these two problems are strictly convex, their minimizers are unique. Next, suppose we would like to estimate θ given a (noiseless) measurement y_i of the minimizer. Observe that FOP-I is identifiable because we must have $\theta = y_i$. However, FOP-II is not identifiable because if $y_i = 1$, then we may have any $\theta \in [1, 2]$. Thus, the constraint $x \leq 1$ renders FOP-II unidentifiable, and precludes the possibility of **IC** holding for FOP-II.

Though FOP-II is not identifiable, a related problem is identifiable because of external inputs. In particular, consider an FOP-III with $x \in \mathbb{R}$ and $\theta \in [0, 2]$ that is given by $\min\{(x - \theta - u)^2 \mid x \leq 1\}$. This problem is strictly convex, and so its minimizer is unique for each fixed value of u . In fact, the minimizer is given by $y_i = \min\{(\theta + u_i), 1\}$. And so a sufficient condition for identifiability of FOP-III is if $\mathbb{P}(u_i \leq -1) > 0$. For instance, if $u_i = -1$ then $y_i = \theta - 1$ and so θ is uniquely determined by y_i . The presence of the input parameter u ensures identifiability of FOP-III.